HLM Investigation of Repeated Measures in Student Performance with Digital Imaging Concepts

ES&CP 9720: Hierarchical Linear Modeling - SP2011 (Osterlind)

Gordon Graber

Introduction

During the summer semester of 2010, a colleague of mine was an intern in the SISLT Introduction to Digital Media 7361/4361 online course, of which I am effectively the primary instructor. This study is a result of a research design which applied an instructional treatment to and collected data for 7361/4361 over three consecutive semesters.

7361/4361 media is a mainly a technical skills development course in which students complete four projects sequentially. They first design and create Web ready images, then audio segments, video segments, and lastly, a Web page project which show cases the previous projects. Students must also engage in online discussion forums related to the projects, as well a written project summary for each. Between fifteen and thirty students enroll in the course each semester, with a mean of number of students of 25.67 over the three semesters in the study. Most are graduate students: over the same period the grad-to-undergrad ratio = 1.66. Graduate students have an extra task for the final project: a usability study. In summer semesters, 7361/4361 is delivered in eight weeks and in spring and fall in sixteen weeks. Having help with the course opened the opportunity to design a research study using students in the course as the subjects. Both of us could share the load of designing the study methods and materials, and we could get a study up and running quickly.

Initially, we had conflicting ideas about what we wanted to study. My colleague wanted to test the effects of concept mapping activities on student performance in the course's projects, and I was interested in looking at the effects of argumentation activities in the discussion forums. As my colleague was there to help me deliver the course for the starting summer semester, I agreed to use their idea for the study. The study would serve as a pilot, applied to one of the course projects; and as a first attempt designing a complete research project on our own.

Relevant Issues and Questions

Of the four media projects required by the course, we chose the digital image project to serve as the vehicle for our investigation. In the digital media project, students produce a set of Web ready images that satisfy a number of criteria. Creating images for the Web involves solving a complex problem in balancing image quality with image file size, and the project lent itself to a concept mapping activity treatment. The factors affecting the resultant quality and file size of a saved Web image can be delineated and the relationships between them quantified in a concept map that would depict the process clearly.

Concept map activities range from students viewing pre-constructed concept maps to constructing their own, and can serve as both learning tools and evaluation tools. Though active concept map construction has been shown to have a stronger affect on learning, we chose to

present a pre-constructed concept map of the factors that students need to consider when saving digital images for use on the Web, due to time and technical constraints. Students would need extra software and instruction were we to require them to construct their own concept map. In an introductory course were many students have novice computing skills, requiring concept map building would have presented and extra burden that would need to be supported. Given the time we had to design the pilot, we opted to have students view a premade concept map of the digital imaging issues.

Many factors affect student learning, and for our study we wanted to control for those that might predict student performance in complex problem solving, such a choosing the proper image format when saving images for the Web. One factor affecting students' work with complex problems is their predisposition for cognitively difficult tasks. Cacioppo and Petty developed a measure of this propensity, the Need for Cognition, at the University of Iowa during the 1980s. The Need for Cognition measure calculates the sum of responses to eighteen survey question, which prompt respondents to rate their attitudes toward cognitive activities. (Example: "I enjoy working on difficult problems.") For the purposes of our study, we believed that the Need for cognition measure would provide an explanation of the variance in student performance.

We also speculated that prior knowledge would also be a factor in student performance. How much each student already knows about creating Web ready digital images would establish a baseline for individual performance. Accordingly, we made provisions to collect a measure of prior knowledge as along with the need for cognition survey at the start of the course.

Our experience with student performance in the digital image project, in past semesters of 7361/4361, is that students make inaccurate decisions when saving image files because they do not understand how the factors which affect image quality and file size are related. Several text based explanations of the factors are presented each semester and it is not clear whether students are learning from these. We theorized that a concept map, visually representing the factors and their relationships explicitly, would enhance student performance in fulfilling the project requirements.

Study design and Methods

The study was devised as an instructional intervention which centered on a partial application of Problem Based Learning (PBL): though the students produced images for the project individually, they were given the role of designers to fulfill a clients request for a series of images in Web compatible image formats. This strategy imparts a context for authentic learning, which is an element of PBL, and helped us to constrain student output over previous semesters, when students were given free reign in the types of images there were allowed to create. Students in our study were required to create a two Web banner images, and one logo image, which employs a limited color set as well as transparency. These requirements would allow us to judge whether the intervention influences student technical skill growth. This measure was contained within the project rubric, and equal to 10 course points.

Three written exams would also provide a second metric to judge performance, and we created a prior knowledge Pretest, an end-of-project Test, and an end-of-semester Posttest. All three tests measured student understanding of the digital image instructional content from a cognitive

dimension. All three exams tested for evidence of learning of digital imaging issues, in between two and five short essay questions. Each of the tests was scored on a 50 point scale.

In addition, a discussion forum activity that prompted students, organized by treatment group, to work through the instructional content collaboratively provided a window into student cognitive growth processes and another measure of learning. The discussion forum would ideally be coded with inter-rater reliability, but for this pilot the number of posts was counted for each student.

For each of the three semesters in the study, students were randomly placed into either a control group or a treatment group, such that graduates and undergraduates were evenly distributed within each. Groups receive the same base project instructions, but receive a different addendum depending on the assigned group: the treatment group received the concept map and the control group received an equivalent text description. Group assignment is transparent to the students, due to the features of the online Learning Management System, Saki, which is the 7361/4361 online course delivery vehicle. Students cannot readily know they have been separated by a group, or which group they are in.

An IRB consent form was supplied, and students were given the opportunity have their data excluded from the study, but aside from that option and the random treatment assignment, the course was delivered in the same fashion to all students. No extra credit was awarded for participation.

Each semester the Pretest and Need for Cognition survey were delivered, along with the IRB consent form, during the first week of the course. The following three weeks were assigned to the digital image project, along with the discussion forum for each of the groups. When the projects have been submitted to the instructor, the end-of-project Test was released. Finally, the Posttest was released at the end of the semester. We also collected student demographic data, including semester, graduate status, gender, and treatment group.

The dependant variables in the study are continuous measures: Pretest, Test, Posttest, and total discussion forum posts (Total Posts). The independent variables are Semester (1,2 or 3); Gender (0 - female, 1- male); Graduate Status (0 - undergraduate, 1- graduate); Treatment (0 - text condition; 1 – concept map condition); and Need for Cognition (continuous, range: -16 - +16).

HLM was chosen as the method of analysis because it allows us to compare individual growth across persons and groups, and can incorporate different measures at different times within a longitudinal study, provided the treatments, measures and covariates are significantly correlated.

Data Collection, Processing and Analysis

Data was collected and stored through Sakai's Tests and Quizzes features, as well as the discussion forum and project assignments. The three written tests, and the need for cognition survey, were graded directly in Sakai. Scoring projects is a manual process, and project scores were not used in the HLM investigation of our pilot study due to time constraints of recoding them. The data collected was prepared using Excel and SPSS for processing with HLM6.

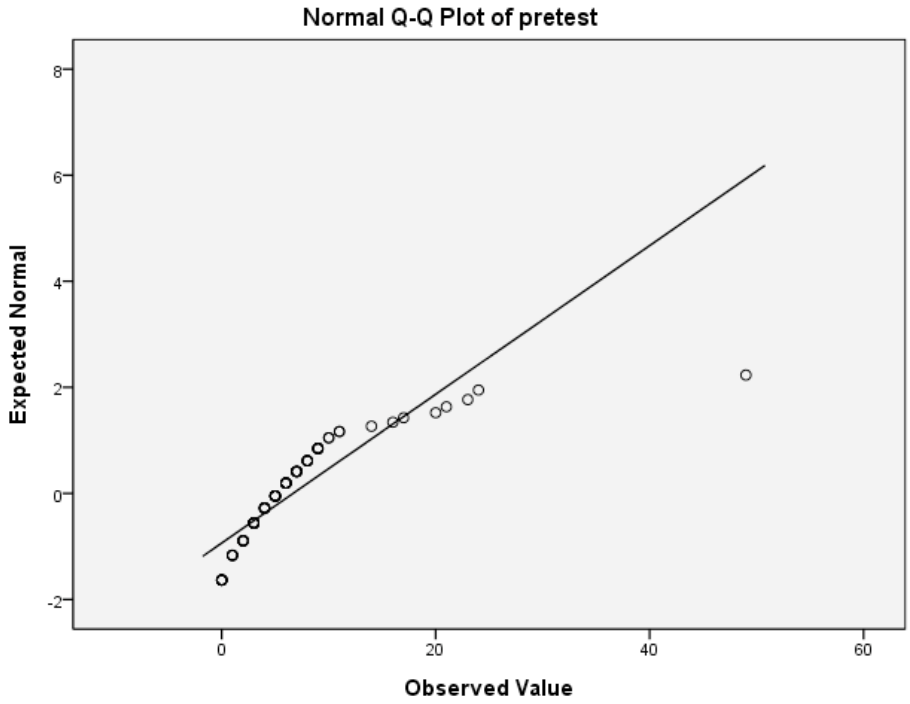
SPSS was used to produce all base and exploratory statistics, including a covariance matrix, which is important in determining whether HLM is a suitable analysis method.

In the complete dataset for three semesters, n = 77 participants, and contained quite a few missing values for some of the measures, which included students who failed to take tests for various reasons. All data fields adhered to roughly normal distributions except for the Pretest, which was positively skewed (3.324) with a large kurtosis. While most students enter the course with little prior knowledge of digital image concepts, often one or two ringers take the course in hopes of padding their grade point average. Rather than delete these outliers, the Pretest scores were transformed using the Log10 function in SPSS. Doing so brought the Pretest scores closer to a normal distribution and kurtosis in line.

Base Statistics and pretest skew:

	semester	gender	treatment	pretest	test	posttest	need for cognition	status	total posts
N Valid	77	77	77	77	61	59	76	77	77
Missing	0	0	0	0	16	18	1	0	0
Mean	2.06	.43	.49	6.68	29.79	32.20	2.84	.62	3.36
Std. Deviation	.848	.498	.503	7.131	12.245	12.345	7.744	.488	3.082
Variance	.719	.248	.253	50.854	149.937	152.389	59.975	.238	9.498
Skewness	-.126	.294	.026	**3.324**	-.490	-.840	.856	-.519	1.071
Std. Error of Skewness	.274	.274	.274	.274	.306	.311	.276	.274	.274
Kurtosis	-1.607	-1.965	-2.053	16.093	-.588	-.095	2.931	-1.777	1.621
Std. Error of Kurtosis	.541	.541	.541	.541	.604	.613	.545	.541	.541

The Q-Q plot of the Pretest scores confirms significant skew.

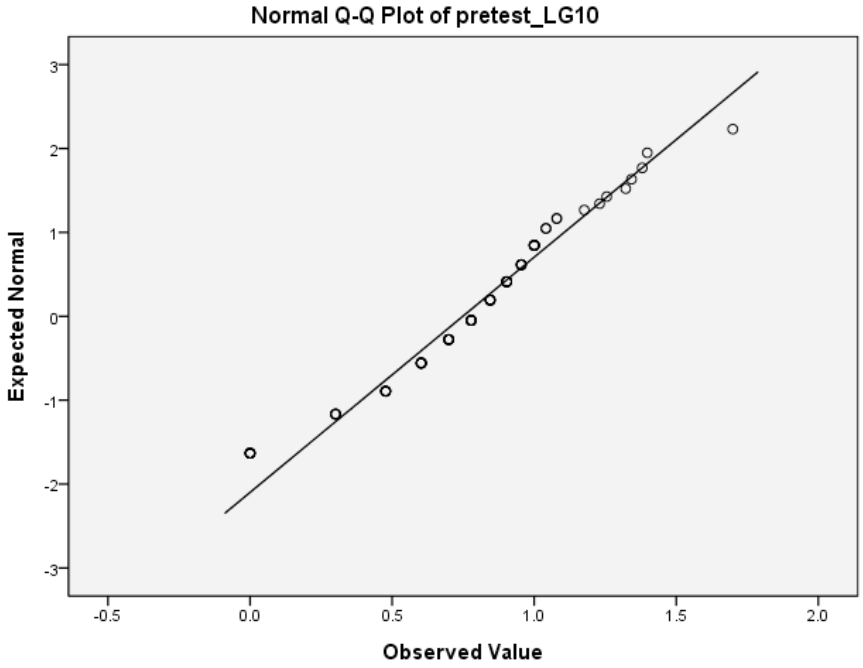


Pretest after Log10 transformation:

pretest_LG10

N	Valid	77
	Missing	0
Mean		.7488
Std. Deviation		.35668
Variance		.127
Skewness		**-.316**
Std. Error of Skewness		.274
Kurtosis		.406
Std. Error of Kurtosis		.541
Range		1.70

The Pretest Q-Q plot after transformation: confirms data has been fitted more closely to a normal distribution:



In preparation for use with HLM6, an aggregate mean of test scores, M_scores, was calculated for each student, and a covariance matrix was created to explore relationships between the data fields. (Note the original non-normal pretest scores included in the covariance matrix in anticipation that the values might be worth examining with HLM.)

An examination of the correlation matrix indicates potential problems with our study design in the lack of significant relationships between our treatment group and measures. Lacking interesting correlations between the dependant and independent variables, any further investigation would not be warranted. However, we can still use HLM to uncover growth trajectories in a longitudinal design that may not be readily apparent.

Fortunately the correlations matrix contains a few highlights, in bold below:

Correlations

		semester	gender	status	treatment	pretest_LG10	pretest	test	posttest	Need Cog	M_scores
semester	Pearson	1	-.036	-.195	.016	.077	-.075	**.528****	.117	.153	.326**
	Sig. (2-tailed)		.759	.090	.887	.503	.518	.000	.379	.188	.004
	N	77	77	77	77	77	77	61	59	76	77
gender	Pearson	-.036	1	-.139	-.015	.025	.117	-.175	.004	-.149	-.191
	Sig. (2-tailed)	.759		.227	.897	.832	.309	.177	.979	.200	.096
	N	77	77	77	77	77	77	61	59	76	77
status	Pearson	-.195	-.139	1	-.144	.209	.165	.031	.300*	.062	.250*

	Sig. (2-tailed)	.090	.227		.211	.068	.152	.813	.021	.593	.029
	N	77	77	77	77	77	77	61	59	76	77
treatment	Pearson	.016	-.015	-.144	1	.115	.163	.179	-.012	-.079	.060
	Sig. (2-tailed)	.887	.897	.211		.317	.158	.168	.930	.496	.607
	N	77	77	77	77	77	77	61	59	76	77
pretest_LG10	Pearson	.077	.025	.209	.115	1	.822**	-.008	.235	-.011	.566**
	Sig. (2-tailed)	.503	.832	.068	.317		.000	.953	.073	.924	.000
	N	77	77	77	77	77	77	61	59	76	77
pretest	Pearson	-.075	.117	.165	.163	.822**	1	-.051	.243	.060	.448**
	Sig. (2-tailed)	.518	.309	.152	.158	.000		.695	.063	.608	.000
	N	77	77	77	77	77	77	61	59	76	77
test	Pearson	.528**	-.175	.031	.179	-.008	-.051	1	.476**	.293*	.778**
	Sig. (2-tailed)	.000	.177	.813	.168	.953	.695		.000	.023	.000
	N	61	61	61	61	61	61	61	56	60	61
posttest	Pearson	.117	.004	.300*	-.012	.235	.243	.476**	1	.084	.859**
	Sig. (2-tailed)	.379	.979	.021	.930	.073	.063	.000		.533	.000
	N	59	59	59	59	59	59	56	59	58	59
needForCog	Pearson	.153	-.149	.062	-.079	-.011	.060	**.293***	.084	1	-.002
	Sig. (2-tailed)	.188	.200	.593	.496	.924	.608	.023	.533		.988
	N	76	76	76	76	76	76	60	58	76	76
M_scores	Pearson	.326**	-.191	.250*	.060	.566**	.448**	.778**	.859**	-.002	1
	Sig. (2-tailed)	.004	.096	.029	.607	.000	.000	.000	.000	.988	
	N	77	77	77	77	77	77	61	59	76	77

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Noteworthy correlations include a relationship between semester and the project Test measure. Between Test and Posttest, and between Test and Need for Cognition. The first indicates some interaction between semester and student performance that is likely due to poor research design (the pretest is not well constructed). The second we would expect, that past performance is related to future performance, and a the .476 correlation between Test and Posttest is a strong indicator that there may be some growth HLM can help explain further. Lastly, Need for Cognition is correlated with the Test, but it is not as significant, only passing at $\alpha = 0.05$. We had assumed Need for Cognition would be an indictor of all the dependant variables, and it is interesting that it only correlates with the Test measure, but not Pretest or Posttest. This may indicate further flaws with the study design. Note other correlations for M_scores, being the mean of Test, Pretest and Posttest, are not interesting in general, because it is accounted for in the individual test scores. Likewise, the large correlation between Pretest_LG10 and Pretest was

expected, because the former is a transformation of the latter. This preliminary analysis of the correlation matrix precedes our HLM investigation with some minor relationships that HLM may help to explain.

The data was then restructured into Level 1 files for use with HLM6. Restructuring employed a measures within persons organization, with each of the Pretest, Test and Post test measures realigned in one column, L1_SCORE, with an added index column, SCOREIND. All other measures were included in the Level 1 file, and in the Level 2 file, for exploration using HLM.

An HLM2 model template as created, specifying the options for measures within persons, and to delete missing cases at runtime. HLM passed the template through initial test, and next the unconditional model for a longitudinal study was created:

LEVEL 1 MODEL

$$L1\_SCORE_{ti} = \pi_{0i} + \pi_{1i}(SCOREIND_{ti}) + e_{ti}$$

LEVEL 2 MODEL

$$\pi_{0i} = \beta_{00} + r_{0i}$$

$$\pi_{1i} = \beta_{10} + r_{1i}$$

where $L1\_SCORE_{ti}$ = the test score for individual i at time t (Pretest, Test, or Posttest). SCOREIND (correspondingly, 1, 2, or 3) was added to index the outcome variable to a given iteration through each test value.

Output from the unconditional model reveals a minor relationship discovered between the cases in the study, with an Intra Class Correlation = 5.26199 / (5.26199 + 126.12776) = 0.04.

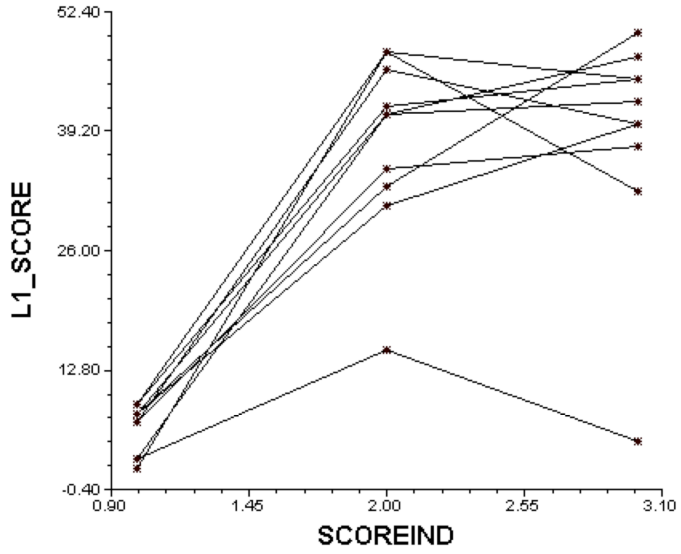
```
Summary of the model specified (in equation format)
 ----------------------------------------------------
Level-1 Model
       Y = P0 + P1*(SCOREIND) + E

Level-2 Model
       P0 = B00 + R0
       P1 = B10

Run-time deletion has reduced the number of level-1 records to 194
Iterations stopped due to small change in likelihood function
******* ITERATION 91 *******

 Sigma_squared =    126.12776


 Tau
 INTRCPT1,P0      5.26199

Tau (as correlations)
 INTRCPT1,P0  1.000
 ----------------------------------------------------
  Random level-1 coefficient   Reliability estimate
 ----------------------------------------------------
  INTRCPT1, P0                        0.095
 ----------------------------------------------------

The value of the likelihood function at iteration 91 = -7.456499E+002
```

The ICC value means that only 4% of the variance detected is across students, which is a very small number. Further, the reliability estimate of 0.095 is tiny, indicating that we have not found anything meaningful. Any relationship reported between students is likely attributable to random error.

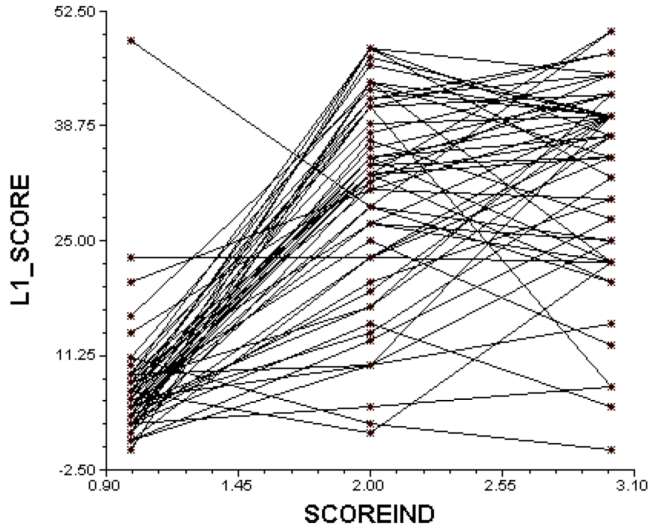
Homogeneity of variance in the test scores had not previously been examined, and it was calculated using HLM6 as a diagnostic check to see if the assumption had been violated. A base model was run with L1_SCORE as the output variable. The results show that we are able to reject the null hypothesis ($p > .5$), and variance in the test scores appears to be randomly distributed:

```
Test of homogeneity of level-1 variance
 ----------------------------------------
 Chi-square statistic          =     57.13979
 Number of degrees of freedom =   62
 P-value                       = >.500
```

Examination of ten test score plots (using the original non-transformed Pretest scores) shows a discernable trend from Pretest to Test – most students appear to improve, but then Posttest scores do not seem to be consistent, with some losing and some gaining over the previous Test value.



A look at all the student test score plots in one graph shows the confusion more clearly:



There seems to be some common trends between students, but there are enough cases where the data trend appears to be unique. It is an indicator the measures are not well constructed.

Having found nothing of value, I wanted to run the empty model with the original non-transformed Pretest scores to see what the difference would be. The unconditional model is the same as the previous run, above. The only difference in this run is that the original Pretest scores have been substituted for the transformed ones. The output shows a much larger ICC (0.146) and stronger reliability estimate (0.339). (Note that a solution was found in just eleven iterations.)

```
Level-1 Model
       Y = P0 + P1*(SCOREIND) + E

Level-2 Model
       P0 = B00 + R0
       P1 = B10

Iterations stopped due to small change in likelihood function
******* ITERATION 11 *******
 Sigma_squared =    117.31269

 Tau
 INTRCPT1,P0     20.05580

Tau (as correlations)
 INTRCPT1,P0  1.000

 ------------------------------------------------------
  Random level-1 coefficient   Reliability estimate
 ------------------------------------------------------
  INTRCPT1, P0                          0.339
 ------------------------------------------------------
The value of the likelihood function at iteration 11 = -6.347845E+002
```

The output shows how sensitive HLM is to data normality, as the original Pretest values used were not normally distributed. If we hand not adjusted for normality, we would have been greatly mislead by the results.

Lastly, a quadratic Level 1 file was considered to see if it would reveal a more robust growth trend. The bends, visible in the plots of the three test scores do not look linear, and a growth curve model appears to be appropriate. The attempt failed due to lack of knowledge about how to create and apply the Level 1 file with squared values.

Conclusions

The preparation of the HLM analysis of this pilot study revealed a number of design and execution flaws with the study. Because my colleague was only involved in the study for the first semester of the three, we did not fully communicate each other's intentions with respect to data coding when they left, much of the data from the first semester had to be recoded. As a result, my ability to process all of the potential data for the purposes of the HLM analysis has been hampered. Lack of inter-rater reliability is also an issue. Data was not consistently scored due to the change of hands, and scoring appears to have evolved over the three semesters data was collected (possibly evidenced by the correlation between semester and Test scores.) The treatment needs to be more explicitly integrated with the measures, so that we can be sure students are using the instructional interventions. Pretest, Test and Posttest measures also need to be redesigned to ensure they measure the performance that the interventions target. All in all the pilot and HLM investigation helped me become more aware of the problems and issues in such a complex study, and better able to redress them in the future.

Despite not finding anything of value using HLM, I believe HLM is suited for this type of complex longitudinal study, and I will be looking forward to learning more about how it can be used to spot growth trends across individuals as evidence of the affects of instructional interventions.