University of California

Los Angeles

A Multilevel Longitudinal Analysis of Teaching Effectiveness

Across Five Years

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

By

Kairong Wang

2013

Abstract of the Thesis

A Multilevel Longitudinal Analysis of Teaching Effectiveness Across Five Years

By

Kairong Wang

Master of Science in Statistics

University of California, Los Angeles, 2013

Professor Hongquan Xu, Chair

Students' rating of a professor's teaching effectiveness is widely endorsed by many schools to evaluate teacher's teaching quality. The purpose of applying this evaluation is to measure teaching effectiveness on a validated tool across teachers and subjects in the same school. It was expected that the ratings can help teachers on their continuous improvement in teaching and instructional decisions. However, do teachers become more effectively with experience growth? Previous studies found a relatively stable trend across time.

The current research studies teaching effectiveness growth trajectory when teaching, to some extent, was related to teachers' performance review. Using five year's continuous data from 6383 courses taught by 1201 teachers from 25 departments in a university in China, this study fits a set of multivariate longitudinal models to examine trends and identify important predictors. Results find that across ten semesters in five years, there is an average increase rate of 0.029 points on a 5 points scale for each semester when all other factors are treated equally. Teachers' teaching experiences, academic ranking, and the class sizes they taught are all valid predictors to predict their teaching effectiveness. Besides, teachers are not growing in the same rate: those who have taught longer appear to have lower growth rate than those who have taught in short of period time. Teachers with higher academic ranks tend to have lower growth rate than those who have lower academic ranks.

The thesis of Kairong Wang is approved.

Jan de Leeuw

Nicolas Christou

Hongquan Xu, Committee Chair

University of California, Los Angeles

2013

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENT

Thank you my dear husband for all your support. Without you, I will never be able to get this done.

Thank you my dear daughters Anabell, Rebecca and Christina. You are the angels of my life and my joy.

Thank you my dear parents, you always stand behind me anywhere and anytime.

Thank you my Committee Chair Hongquan Xu, for being such a wonderful professor and a great mentor during this whole process. Thank you so much for guiding me to get the thesis done.

Thank you professor Nicolas Christou, for being so nice and kind. I really enjoy your class and your talk.

Thank you Professor Jan de Leeuw, for the wonderful lessons you have taught and for all the help you gave me.

Thank you my dear friend Dr. Hong Wei. Thanks for the data and all the help.

Thank you dear Glenda Jones, for all your help and friendship.

CHAPTER 1

Introduction

Students' rating on teaching effectiveness of teachers is widely endorsed by many schools in order to evaluate their teaching quality. It is expected that this type of evaluation provides constructive feedback to the teachers and help them make a difference in their teaching achievements. By using the regularly collected evaluation data, teachers are able to acknowledge their strengths and weaknesses. Eventually, these results can help them enhance their teaching (McKeachie, 1997). Meanwhile, in some universities, the student rating data are saved in teacher's profile and are used as a primary indicator for promotion and performance reviews (McKeachie, 1997; Pike, 1998).

There were numbers of studies about students' evaluation on teaching effectiveness in the past twenty years. These studies fall to two categories: cross-sectional studies and longitudinal studies.

Cross-sectional studies mainly focused on investigating some important variables, such as: class size, students' grade, and course format (Otani, et al., 2011). Researchers are interested in describing the mean difference across groups. For example, in a sex difference study, Halpern (2000) found that the men and women are different in their distribution across the disciplines. Researchers also found that, generally, men receive higher ratings than women and white faculty tends to receive higher ratings than minority faculty. (Nettles et al., 2000). Moreover, Marsh

(1981) examined the relative influence between course and teacher and found that teachers' teaching effectiveness was primarily a function of teacher who teaches a course rather than the course that is taught. These kinds of studies suggest that individual difference in teaching is due to some uncontrollable factors.

Longitudinal studies mainly focused on identifying the teachers' teaching effectiveness growth trajectory. Mash (1980) collected data from the same students at different time point to evaluate the teachers' teaching effectiveness. The first time point is when the student attending the course. The second time point is one year after the students graduated. Results showed that students used the same perspectives to evaluate their teacher at the two time points. The ruler they used to evaluate the teacher did not change over the time even after students apply the course materials in further course work or after graduation. Marsh (2007) also examined 13 years teaching effectiveness data by applying multilevel models. This study showed that there was little evidence that teachers became either more or less effective with added experience. By using the same method, Carle (2009) analyzed the data from 10,392 classes which were taught by 1120 teachers across three years and found that students rating of teaching effectiveness remained relatively stable across time and different teachers. Other factors (such as discipline, course level, set, minority status and tenure) have no significantly effects on teaching effectiveness.

In the longitudinal studies summarized above, previous researchers did not mention how the ratings from each evaluation were applied on the teachers' future teaching and how these ratings were directly related to teachers' benefits or performance reviews. In this study, teachers were under the pressure that their courses may become an open teaching course. Then their courses will be evaluated not only by students but also by educational experts. This study will examine the individual teacher's growth trend under teaching pressure and some predictors that were collected at the same time.

CHAPTER 2

Data and Analysis

2.1 Data Source

Data used in this study are from the Center for Teaching and Learning in Beijing Normal University. The Center for Teaching and Learning dedicates to supporting and enhancing teaching and learning for the whole university. The center assesses teachers' teaching effectiveness at the end of each semester by using a validated questionnaire named "Students' Evaluation of Teaching Effectiveness". At the end of each semester when the students' final tests are finished, employees from the center visit the class and ask the students to answer the questionnaire. Collected questionnaires will then be sent back to the center for further analyses. The questionnaire includes 23 items. Twenty-two out of the 23 items are rated by Likert scales from 1 to 5 with 1 stands for poor and 5 stands for excellent. The last question is an open-ended question asking for students suggestions. Of the 22 Likert scale questions, one is about the overall rating toward the teacher's teaching effectiveness. A class mean based on students who attended the class is recorded as the indicator of the teachers teaching effectiveness level.

2.2 Sample

The five years data set includes 6383 courses taught by 1201 teachers from 25 departments. There are 2528 distinct courses in all. Some teachers teach more than one course in each semester. Some teachers teach two different courses in two consecutive semesters. Although the best method to analyze this type of data is to take course as an independent variable, it is difficult for this analysis to do so because of too much nonconsecutive data. Therefore, teachers' teaching scores for different courses are aggregated in each semester. This way, teachers will have one rating score for one semester even though they may teach one or more courses. Table 2.1 lists the teacher frequency in each semesters. 13% of the teachers taught all 10 semesters and 15% of the teachers taught only 1 semester.

Semesters	Teacher Numbers	Proportion
1	186	15%
2	65	5%
3	106	9%
4	124	10%
5	159	13%
6	116	10%
7	103	9%
8	94	8%
9	92	8%
10	156	13%
Total	1201	100%

Table 2.1 Sample Distribution across Ten Semesters

2.3 Data Description

Outcome Variable:

The outcome variable of this study is measured by asking a single question of "What is your overall evaluation toward the teacher who taught the class that you are attending this semester?" in a 5 point Likert scale with 1 represents poor and 5 represents excellent.

Independent variables:

Teacher Academic Ranking: The school archive has the information of each teacher's academic ranks. The coding of this rank in the study is listed in table 2.2.

Code	Teacher Numbers	Proportion
1	Professor	29%
2	Associate Professor	39%
3	Assistant Professor	22%
4	Lecturer	1%
5	Teacher (all other ranks)	9%
Total	1201	100%

Table 2.2: Sample Distribution across Teachers Academic Ranking

Course Type: There are two types of courses taught in this university. General courses with coding 1 are those that students from all departments can take. Major courses with coding 0 are those that are taught only for the students who are in this major. There are 89% of major courses and 11% of general courses taught in the five years from 2004 to 2008.

Teacher Service Years: the amount of years that the teacher taught since he / she started to work. This variable was calculated by using the year that the teacher was first evaluate from 2004 to 2008 minus the year that this teacher started to work. Teacher Service Years has a minimum value of 1 and maximum value of 49. The median value for this variable is 20.

Class Size: Class size is measuring how many students are in the class that the teacher is teaching. If a teacher teaches more than one courses at a particular semester, the mean of student numbers for all the classes that were taught by the same teacher will be used. This variable has a minimum value of 2 and a maximum value of 227. The median value for this variable is 37.

2.4 Research Questions

Did the Beijing Normal Universities teachers' teaching effectiveness improved from 2004 to 2008? If so, what is the growth trend? Did all the teachers share the same growth trend? If not, what is the difference among the teachers? What are the important predictors that can be used to predict the teachers' future teaching effectiveness? How do these predictors work together to decide a teacher's teaching effectiveness?

2.5 Data Analysis Methods

The dataset in this research is longitudinal. Data was repeatedly collected at 10 fixed time points. Based on previous research, two types of models that are widely used to treat longitudinal data may be adopted.

2.5.1 Univariate or Multivariate Analysis of Variance (ANOVA and MANOVA)

ANOVA and MNOVA are widely used method to deal with variance analysis. The purpose of this type of variance analysis is to test the significant difference between several groups of means. The methods assume equal level of variance is existed across groups. It is not informative about individual growth.

2.5.2 Multilevel Analysis of Longitudinal Data

When the analysis interest is going beyond group difference and also cares about the individual change over time, a multilevel model is needed. As pointed by Singer (2003), a two level multilevel analysis includes two components: (1) a level 1 submodel at repeated time level which inspects how individuals change over time; and (2) a level 2 submodel at subject level which examines individual difference of change across time.

The two components can be represented by mathematical expressions. Using the notations from Brky & Raudenbush (2002), the two levels of components would be:

Level 1 (repeated measures): $$y_{ti} = \pi_{0i} + \pi_{1i} Time_{ti} + e_{ti}$$

Level 2 (subject level): $$\pi_{0i} = \beta_{00} + \gamma_{00i}$$

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

Where t is time and i is the individual sample in the data. y_{ti} represents the estimated outcome measures for individual i at the time of t. π_{0i} is the intercept of the regression and π_{1i} is the slope. e_{ti} is the unestimated residual indicating the variability of the data around the regression line. In the current analysis, teacher 1's intercept is π_{01} , teacher 2's intercept is π_{02}. π_{01} represents teacher 1's true value at time of 0. π_{11} represents the slope of teacher 1 change trajectory. If π_{11} is positive, teacher 1 's true outcome will increase across time; if π_{11} is negative, teacher 1's true outcome will decrease over time. β_{00} is the population average of true initial status. $\gamma_{00i}$is the variance between individual intercept and the average intercept. β_{10} indexes the population average change of rate. γ_{10i} is the variability of individual change rate around the average population change rate. Combining the level 1 and level 2 models, we can have a single regression model represents the individual change over time:

$$y_{ti} = \pi_{0i} + \pi_{1i}Time_{ti} + e_{ti} = \beta_{00} + \gamma_{00i} + (\beta_{10} + \gamma_{10i})Time_{ti} + e_{ti}$$
$$= \beta_{00} + \beta_{10}\, Time_{ti} + \gamma_{00i} + \gamma_{10i}Time_{ti} + e_{ti}$$

When additional predictors are introduced, the above models can be extended as:

Level 1 (repeated measures): $y_{ti} = \pi_{0i} + \pi_{1i}Time_{ti} + \pi_{2i}\, X_{ti} + e_{ti}$

Level 2 (subject level): $\pi_{0i} = \beta_{00} + \beta_{01}Z_i + \gamma_{00i}$

$$\pi_{1i} = \beta_{10} + \beta_{11}Z_i + \gamma_{10i}$$

$$\pi_{2i} = \beta_{20} + \beta_{21}Z_i + \gamma_{20i}$$

Where X is time varying variable and Z is time invariant variable.

A single equation from above can be written as:

$$y_{ti} = \beta_{00} + \beta_{10}\, Time_{ti} + \beta_{20}X_{ti} + \beta_{01}Z_i + \beta_{11}Time_{ti}Z_i + \beta_{21}Time_{ti}Z_i + \gamma_{10i}Time_{ti}$$
$$+ \gamma_{20i}X\_ti + \gamma_{00i} + e_{ti}$$

For the dataset in this analysis, the two levels of analysis would be:

Level 1 (repeated measures): $y_{ti} = \pi_{0i} + \pi_{1i}Time_{ti} + e_{ti}$

Level 2 (subject level):

$$\pi_{0i} = \beta_{00} + \beta_{01}Classsize_i + \beta_{02}serviceyear_i + \beta_{03}academicrank_{ti} + \beta_{04}Coursetype_{ti} + \gamma_{00i}$$

$$\pi_{1i} = \beta_{10} + \beta_{11}Classsize_i + \beta_{12}serviceyear_i + \beta_{13}academicrank_{ti} + \beta_{14}Coursetype_{ti} + \gamma_{10i}$$

A single equation from above for this research is expressed as:

$$y_{ti} = \beta_{00} + \beta_{01}Classsize_i + \beta_{02}serviceyear_i + \beta_{03}academicrank_{ti} + \beta_{04}Coursetype_{ti} + \beta 10\,Time_{ti} + \beta_{11}Classsize_iTime_{ti} + \beta_{12}serviceyear_iTime_{ti} + \beta_{13}academicrank_{ti}Time_{ti} + \beta_{14}Coursetype_{ti}Time_{ti} + \gamma_{10i}Time_{ti} + \gamma_{00i} + e_{ti}$$

CHAPTER 3

Results

3.1 Descriptive Analysis of the Teaching Effectiveness Data

3.1.1 Mean and Standard Deviation

Overall descriptive statistics is presented in table 3.1.

Mean	Std Dev	Min	1st Qu	Median	3rd Qu	Max
4.46	0.29	2.12	4.31	4.52	4.69	5.00

Table 3.1: The Summary of Teaching Effectiveness Scores

The teaching effectiveness scores from year of 2004 to 2009 (ten semesters) are depicted in the box plot in Figure 3.1. A growth trend over time is shown from this graph. This same trend is shown from the scatter plot (Figure 3.2) in which the relationship of average teaching effectiveness scores and time are examined. The pattern in Figure 3.2 shows that the teaching effectiveness growth level between two semesters is lower than growth level between two consecutive years. Although a clear linear relationship existed between time and scores, we cannot draw a conclusion regarding this relationship because the data is longitudinal data and most teachers are not rated continuously in every semester.

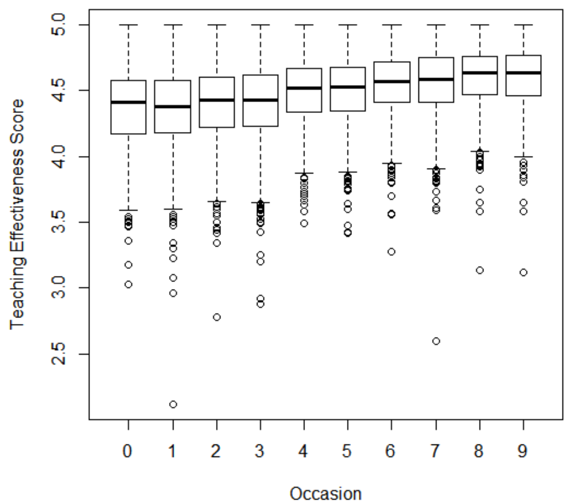

Figure 3.1: The Box-Plot of Teaching Effectiveness Scores across Ten Semesters

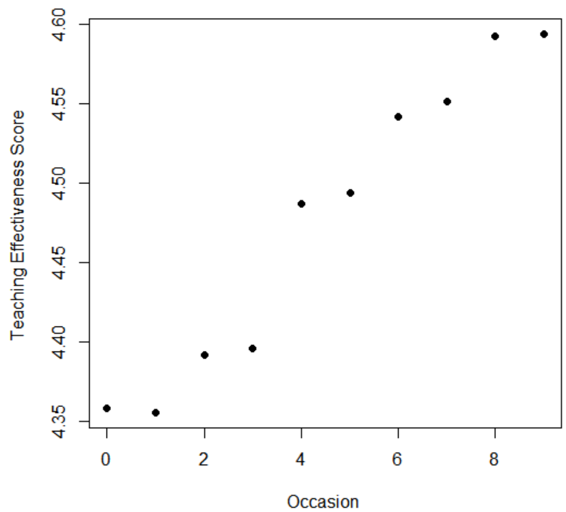

Figure 3.2 The Scatter Plot of the Mean Scores for the Ten Semesters

3.1.2 Individual Growth Pattern

In order to examine the growth pattern at individual level, nine teachers are randomly selected from the data. The growth trajectories for the nine teachers are plotted in Figure 3.3.

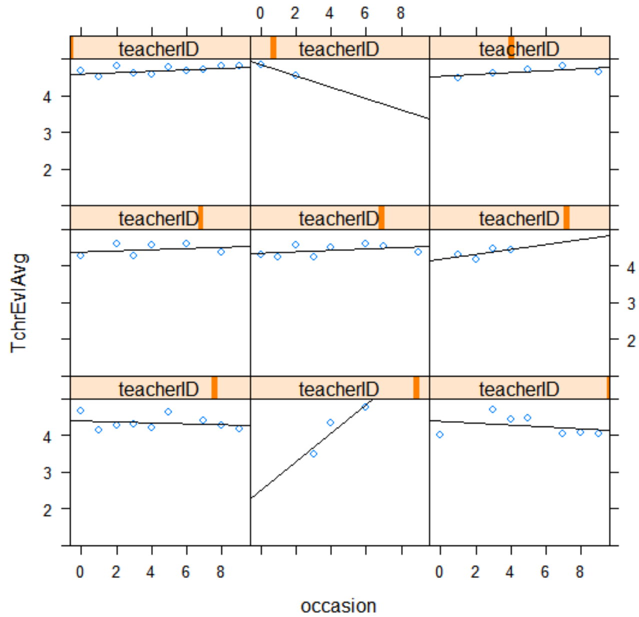

Figure 3.3: Empirical Growth Plots with Superimposed OLS trajectories for 9 Random Selected Teachers

As shown in the figure 3.3, it appears that there is linear relationship between Teaching Effectiveness and Time. This suggests that a linear model could possibly be built. Also, the nine teachers seem to have different growth patterns: some have growing and declining trends and some have not changed over the years. This is an indication that a multilevel analysis should be considered.

To identify potential predictors, plots of OLS fitted trajectories are separately created by levels for selected variables. As shown in Figure 3.4, the growth trend for general course and major course appears the same. Larger classes seem to have higher growth rates than small classes. Those who are classified as teacher in the academic rank appear to have lower starting points but higher growth rate compared to professors. Moreover, there is no difference between the trend for teachers with less teaching experience and those who have rich experience. These figures suggest the initial idea of how the predictor should be included in the study.

We can see that there is individual different growth trajectory from the above analyses. Multivariate Analysis of Variance (MANOVA) is not a good option to analyze this research data based on the discussion in Chapter 2. A better method will be to conduct a multilevel longitudinal study that can deal with incomplete data and can analyze the data at individual level.

3.2 Multilevel Longitudinal Analyses

A set of two-level (level 2 = teacher, level 1= time) growth models are fitted to examine teachers teaching effectiveness growth with time and to identify important variables which can predict the change rate of teaching effectiveness. Maximum likelihood estimation method in SAS Proc Mixed (SAS 9.3) is used to conduct all the modeling analyses.

In the following analyses, teachers' teaching effectiveness score is the outcome variable. Predictor variables include: time (0 = 1st semester in 2004, 1 = 2nd semester in 2004 ... 8 = 1st semester in 2008, 9 = 2nd semester in 2008), teacher's academic rank (1 = professor, 2 = associate professor, 3 = assistant professor, 4 = lecturer, 5 = teacher), course type (0 = major course, 1 = general course), class size (number of students in the class), and service years (years of teaching experiences at the first measurement).

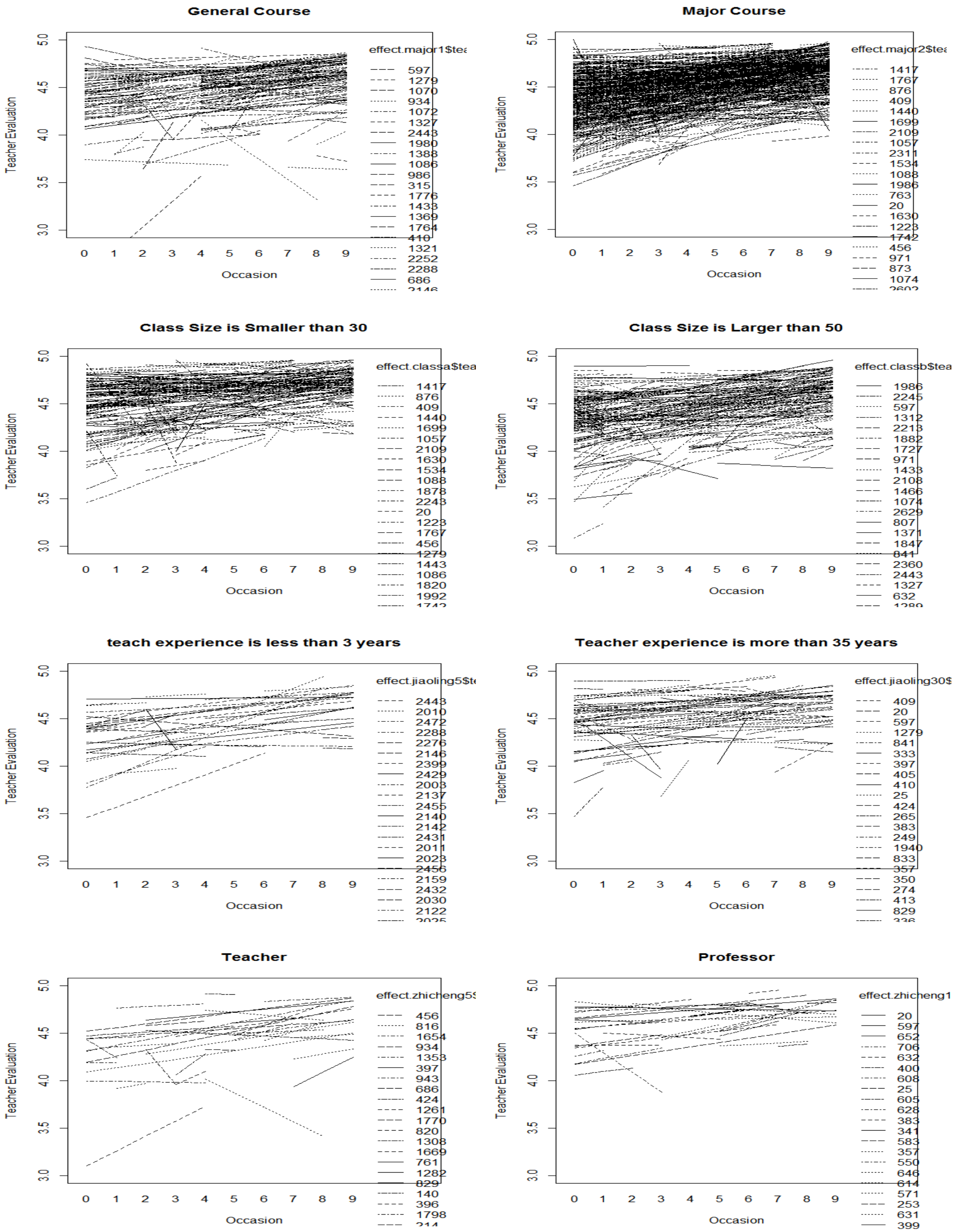

Figure 3.4 Predictor variables change with Time at Different Levels

Model 1: Multilevel models are often fitted starting from an unconditional means model which has no predictors included. This model only includes intercept terms and the partitions of the outcome variances. It serves as a baseline model to be compared with the future models (Singer & Willett, 2003). The average teaching effectiveness score is 4.462, which is the intercept of the unconditional means model. Variance is decomposed to two parts: Level 1 variance is the repeated measure across time within individual and level 2 variance is the part between individuals. About 47% ($\frac{0.41}{0.41+0.47}$) of the variance is explained by the individual difference.

Model 2: Model 2 is the unconditional growth model where the fixed and random linear of time effects are added. The fixed effects for the initial point (intercept) and the slope of average population change trajectory for teaching effectiveness are both significant. Entering the linear time effect decreases the level 1 variance by 23%.

$$R\epsilon^2 = \frac{\sigma_\epsilon^2(model\ 2) - \sigma_\epsilon^2(model\ 1)}{\sigma_\epsilon^2(model\ 2)} = \frac{0.47 - 0.36}{0.47} = 0.23$$

We conclude that 23% of the within-person variation in teaching effectiveness is explained by linear time. Since the remaining time related level 1 variance (δ_ε^2) is still significant, more time varying predictors should be introduced to level 1 submodel. The level 2 variance components are associated with the individual teacher growth parameters. The significant initial status (δ_0^2) variance means that individuals have different initial status. It also indicates that more explanatory factors should be added for further analyses. The significant variance of rate of change (δ_1^2) means individual teachers have different of rate of change with time. The covariance residual (δ_{01}) assess the relationship between the initial status and change rate. Correlation coefficient of the relationship can be calculated based on these variance components from the formula provided by Singer and Willett (2003):

$$\rho_{01} = \frac{\sigma_{01}}{\sqrt{\sigma_0{}^2\sigma_1{}^2}} = -\frac{0.003}{\sqrt{(0.065)(0.0004)}} = -0.59$$

Therefore, there is a moderate negative relationship between the initial status and the changing rate. Teachers who have lower starting points tend to have higher growth rate.

Model 3: As suggested by Model 2, more explanatory factors are introduced. These factors include time invariant factors of teacher academic ranking, major course, service years at the first measurement, and the time varying variable class size. These variables are put under the fixed effects part. As shown in table 3.2, the two variables of teacher academic rank and class size have significant negative relationship with teaching effectiveness. For the teacher academic rank variable (1 means professor and 5 means teacher), the lower value stands for higher rank. Therefore, the negative relationship indicates that higher rank teachers have higher average teaching effectiveness. Same for the class size variable: the larger class size that the teacher teaches, the lower teaching effectiveness score the teacher tends to receive. Both course type (major course or general course) and service years have no significant effects on teaching effectiveness.

Adding these explanatory variables slightly help to decrease level 1 within-person variance. It also helps reduce the level 2 teachers' individual difference variance from 0.65 to 0.61. These four variables explain 6% of the between teacher variance.

Comparing the Model fit indices AIC, BIC and deviance between model 2 and model 3 suggests that adding the new variables improves model fit.

Model 4: The interaction term between time and the new added explanatory variables from model 3 are included in model 4. The final model exclude the variable of course type since there is no significant relationship between this term and the teaching effectiveness. Both are smaller than in Model 3 when the fit indices AIC and BIC are inspected. Therefore, model 4 fits the data better and it is appropriate to keep model 4 as the final model.

3.3 Interpretation of the Terms in the Final Model

For the fixed effects:

Time (0.029): When all other variables are equal, the average of teaching effectiveness scores increase 0.029 for every semester.

Teacher Academic Rank (-0.043): Teachers' teaching effectiveness scores appear to be growing 0.043 when teachers are promoted one academic rank higher (e.g., from teacher to instructor or from associate professor to professor), on average, when all other variables are equal.

Class Size (-0.001): Controlling for the effects of other variables, the estimated teaching effectiveness score for this teacher appear to drop 0.001 on average when each additional student is added in the class,

Service Year (0.003): Controlling for the effects of other variables, the teachers appear to receive 0.003 teaching effectiveness score points on average when they serve one more teaching year.

Interaction between Time and Teacher Academic Rank (0.003): For teachers moving one rank higher (e.g., from associate professor to professor) when all other variables are equal, the average teaching effectiveness score rate of change with time is 0.003 lower, meaning that the higher academic rank teachers have a lower score growth rate with time.

Interaction between Time and Service Year (-0.003): Controlling for the effect of other variables, for 1 more year of teaching service, the average rate of change in teaching effectiveness is 0.003 lower. It indicates that the longer time the a teacher teaches, the lower growth rate he / she will receive.

3.4 Normality Check of the Residual

Normal Q-Q plot (Figure 3.5) and the scatterplot (Figure 3.6) for the standardized residuals are used to check the normality of the residuals. The normal Q-Q plot appears approximately linear and the standardized residual by ID scatter plot also appear to conform to normal theory assumption. A large of the standardized residuals is in the interval of above and below 2 standard deviations of center. There are only a small proportion of the values deviated beyond.

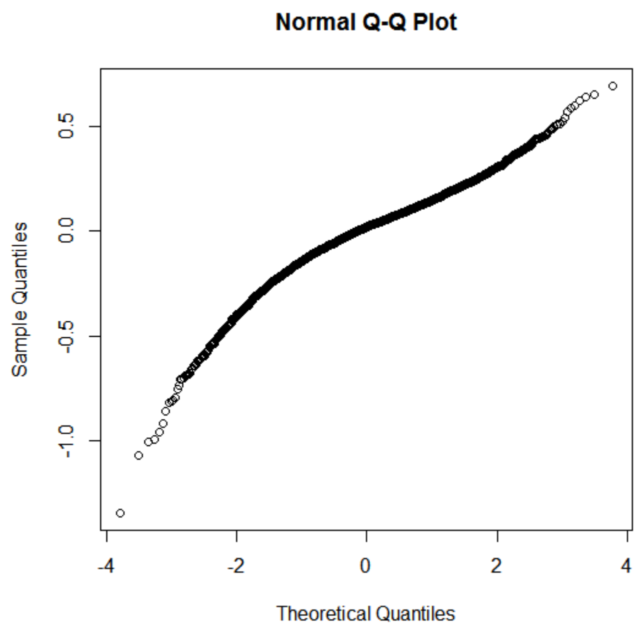

Figure 3.5 Q-Q Plot of the Final Model Residuals

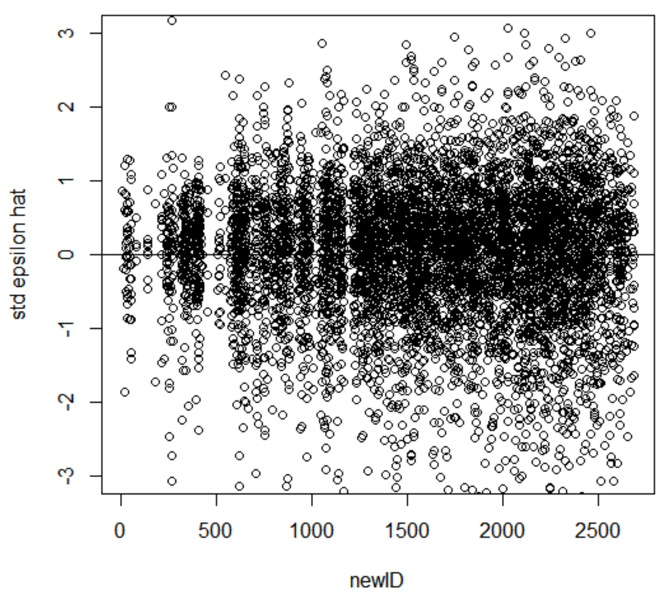

Figure 3.6 Standardized Residuals of the Final Model

		Model 1	Model 2	Model 3	Model 4
Fixed Part		Coeff. (s.e.)	Coeff. (s.e.)	Coeff. (s.e.)	Coeff. (s.e.)
intercept		4.462*** (0.0069)	4.322*** (0.0092)	4.416*** (0.0237)	4.423*** (0.0317)
Teacher Academic Rank				-0.025*** (0.0060)	-0.043*** (0.0085)
Course Type				0.015 (0.0101)	
Class Size				-0.001*** (0.0001)	-0.001*** (0.0001)
Service Years				0.001 (0.0001)	0.003*** (0.0010)
Rate of Change					
Time			0.031*** (0.0011)	0.031*** (0.0011)	0.029*** (0.0038)
Time* Teacher Academic Rank					0.003*** (0.0010)
Time * Service Years					-0.0003*** (0.0001)
Variance Components					
Within-person	Level 1: δ_ε^2	0.047*** (0.0009)	0.036*** (0.0008)	0.035*** (0.0008)	0.035*** (0.0008)
In initial status	Level 2: δ_0^2	0.041*** (0.0022)	0.065*** (0.0040)	0.061*** (0.0038)	0.060*** (0.0039)
In rate of change	Level 2: δ_1^2		0.0004*** (0.0000)	0.0004*** (0.0001)	0.0003*** (0.0001)
Covariance	Level 2: δ_{01}		-0.003*** (0.0004)	-0.003*** (0.0004)	-0.003*** (0.0004)
Fit Indices					
Deviance		466.5	-671.0	-848.2	-873.6
AIC		472.5	-659.0	-828.2	-851.6
BIC		487.5	-629.0	-778.3	-796.7

*** P<0.01

Table 3.2 Results of Fitting a Taxonomy of Multilevel Models for Change to the Teaching Effectiveness

CHAPTER 4

Conclusion

Multilevel longitudinal analyses using the dataset in this research show that Beijing Normal University teachers teaching became more effective from 2004 to 2008. Across the ten semesters in five years, there is an average increase rate of 0.029 points in a 5 point scale for each semester when all other factors are treated equally. Teachers' teaching experiences, academic ranking, and the class sizes they taught are all valid predictors to predict their teaching effectiveness. Teachers are not growing in the same rate: those who have taught longer appear to have lower growth rate than those who have taught in short of period time. High academic rank teachers tend to have lower growth rates than those who have lower academic ranks. Residual components analyses from the final model indicate that there are some other important factors which can help explain the teachers teaching effectiveness change. Further analyses in this topic are needed to find more important variables.

Studies on teaching effectiveness conducted by Marsh (2007) and Carle (2009) revealed that teachers' teaching effectiveness remained relatively stable across time and teachers. Results drawn from this study is controversy with their studies. This difference can be examined by analysis on how the teaching effectiveness scores were applied to future teaching for teachers.

In Marsh and Carle's studies, teachers are evaluated by the students every semester or school year. Teachers were provided with the results once the ratings are scored. There is no further application of these ratings required by the university. Teachers voluntarily decide whether they would like to make changes according to these feedbacks. However, these rating for the teachers in Beijing Normal University acted a different role. There was a university wide education reform activity going on in Beijing Normal University from 2004 to 2008. The university was under evaluation to be nominated as an extraordinary universities of China. Teachers had the pressure that their teaching maybe assigned to be evaluated not only by the students, but also by the education experts from the nomination committee. Numerous teachers used the ratings as feedback to help them to get prepared for their possible public class evaluation. Therefore, the growth trend from 2004 to 2008 for the teachers in Beijing Normal University could possibly be due to these teachers subjectively desire to improve on effectiveness in teaching. Further research should consider an education experimental design in which control groups and treatment groups are assigned. Considering the important predictors from this research and previous research, longitudinal data can be collected from factorial design studies.

Appendix: SAS Results

Model 1 Detailed Summary

The Mixed Procedure

Model Information	
Data Set	MYWORK.FINALDATA
Dependent Variable	TchrEvlAvg
Covariance Structure	Unstructured
Subject Effect	teacherID
Estimation Method	ML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Dimensions	
Covariance Parameters	2
Columns in X	1
Columns in Z Per Subject	1
Subjects	1102
Max Obs Per Subject	10

Number of Observations	
Number of Observations Read	6383
Number of Observations Used	6383
Number of Observations Not Used	0

Iteration History

Iteration	Evaluations	-2 Log Like	Criterion
0	1	2428.84582218	
1	2	466.70687797	0.00004220
2	1	466.46267109	0.00000008

Iteration History			
Iteration	**Evaluations**	**-2 Log Like**	**Criterion**
3	1	466.46223686	0.00000000

Convergence criteria met.

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr > Z**
UN(1,1)	**teacherID**	0.04107	0.002248	18.27	<.0001
Residual		0.04694	0.000916	51.24	<.0001

Fit Statistics	
-2 Log Likelihood	466.5
AIC (smaller is better)	472.5
AICC (smaller is better)	472.5
BIC (smaller is better)	487.5

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
1	1962.38	<.0001

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	4.4619	0.006854	1101	651.03	<.0001

Model 2 Detailed Summaries

The Mixed Procedure

Model Information	
Data Set	MYWORK.FINALDATA
Dependent Variable	TchrEvlAvg
Covariance Structure	Unstructured
Subject Effect	teacherID
Estimation Method	ML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Dimensions	
Covariance Parameters	4
Columns in X	2
Columns in Z Per Subject	2
Subjects	1102
Max Obs Per Subject	10

Number of Observations	
Number of Observations Read	6383
Number of Observations Used	6383
Number of Observations Not Used	0

Iteration History

Iteration	Evaluations	-2 Log Like	Criterion
0	1	1845.46807432	
1	2	-670.59283649	0.00006288
2	1	-670.99614752	0.00000018
3	1	-670.99729490	0.00000000

Convergence criteria met.

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	**teacherID**	0.06452	0.004004	16.11	<.0001
UN(2,1)	**teacherID**	-0.00335	0.000413	-8.12	<.0001
UN(2,2)	**teacherID**	0.000368	0.000057	6.44	<.0001
Residual		0.03587	0.000770	46.58	<.0001

Fit Statistics	
-2 Log Likelihood	-671.0
AIC (smaller is better)	-659.0
AICC (smaller is better)	-659.0
BIC (smaller is better)	-629.0

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
3	2516.47	<.0001

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	4.3217	0.009227	1101	468.36	<.0001
occasion	0.03128	0.001100	1014	28.43	<.0001

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
occasion	1	1014	808.44	<.0001

Model 3 Detailed Summaries

The Mixed Procedure

Model Information	
Data Set	MYWORK.FINALDATA
Dependent Variable	TchrEvlAvg
Covariance Structure	Unstructured
Subject Effect	teacherID
Estimation Method	ML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Number of Observations	
Number of Observations Read	6383
Number of Observations Used	6355
Number of Observations Not Used	28

Iteration History

Iteration	Evaluations	-2 Log Like	Criterion
0	1	1521.01702496	
1	2	-847.41002997	0.00012122
2	1	-848.20512786	0.00000069
3	1	-848.20947576	0.00000000

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
UN(1,1)	**teacherID**	0.06065	0.003846	15.77	<.0001
UN(2,1)	**teacherID**	-0.00323	0.000403	-8.01	<.0001
UN(2,2)	**teacherID**	0.000372	0.000057	6.58	<.0001

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
Residual		0.03499	0.000754	46.43	<.0001

Fit Statistics	
-2 Log Likelihood	-848.2
AIC (smaller is better)	-828.2
AICC (smaller is better)	-828.2
BIC (smaller is better)	-778.3

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
3	2369.23	<.0001

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	4.4158	0.02367	1089	186.52	<.0001
occasion	0.03065	0.001098	1008	27.90	<.0001
zhichengID	-0.02474	0.005976	4252	-4.14	<.0001
majorclass	0.01459	0.01005	4252	1.45	0.1465
classizeAvg	-0.00100	0.000087	4252	-11.50	<.0001
jiaoling	0.000794	0.000678	4252	1.17	0.2419

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
occasion	1	1008	778.52	<.0001
zhichengID	1	4252	17.14	<.0001
majorclass	1	4252	2.11	0.1465
classizeAvg	1	4252	132.29	<.0001
jiaoling	1	4252	1.37	0.2419

Model 4 Detailed Summaries

The Mixed Procedure

Model Information	
Data Set	MYWORK.FINALDATA
Dependent Variable	TchrEvlAvg
Covariance Structure	Unstructured
Subject Effect	teacherID
Estimation Method	ML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Number of Observations	
Number of Observations Read	6383
Number of Observations Used	6355
Number of Observations Not Used	28

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
UN(1,1)	**teacherID**	0.05961	0.003767	15.82	<.0001
UN(2,1)	**teacherID**	-0.00305	0.000391	-7.82	<.0001
UN(2,2)	**teacherID**	0.000341	0.000055	6.22	<.0001
Residual		0.03502	0.000754	46.46	<.0001

Fit Statistics	
-2 Log Likelihood	-873.6
AIC (smaller is better)	-851.6
AICC (smaller is better)	-851.6

Fit Statistics	
BIC (smaller is better)	-796.7

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
3	2372.43	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	4.4226	0.03165	1089	139.72	<.0001
occasion	0.02915	0.003832	1006	7.61	<.0001
zhichengID	-0.04273	0.008539	4253	-5.00	<.0001
classizeAvg	-0.00098	0.000086	4253	-11.46	<.0001
jiaoling	0.002674	0.000944	4253	2.83	0.0047
occasion*zhichengID	0.003340	0.001079	4253	3.09	0.0020
occasion*jiaoling	-0.00033	0.000116	4253	-2.85	0.0044

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
occasion	1	1006	57.88	<.0001
zhichengID	1	4253	25.04	<.0001
classizeAvg	1	4253	131.23	<.0001
jiaoling	1	4253	8.02	0.0047
occasion*zhichengID	1	4253	9.57	0.0020
occasion*jiaoling	1	4253	8.11	0.0044

REFERENCES

Halpern, D. F. (2000). Sex Differences in Cognitive Abilities (3rd Edition). Mahwah, NJ: Lawrence Erlbaum, Associates, Inc. Publishers.

Hox, Joop (2010). Multilevel Analysis: Techniques and Applications, Second Edition, Routledge Academic(2nd edition).

Marsh, H. W. (2007b). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. Journal of Educational Psychology, 99, 775–790.

Marsh, H. W. (1980). The influence of student, course and instructor characteristics on evaluation of university teaching. American Educational Research Journal, 17, 219-237.

Marsh, H. W. (1981). The use of path analysis to estimate teacher and course effects in student rating of instructional effectiveness. Applied Psychological Measurement, 6, 47–60.

McKeachie, W. J. (1997). Student ratings: The validity of use. American Psychologist, 52, 1218–1225.

Nettles, M. T., Perna, I.. W.. &Braddburn, E.M. (2000). Salary, Promotion, and Tenure Status of Minority and Women Faculty in U.S. Colleges and Universities. Education Statistics Quartely, 2, 94-96.

Otani, K., Kim, B.J., and Cho, J.I.(2012). Student evaluation of teaching (SET) in higher education: How to use SET more effectively and efficiently in Public Affairs Education. Journal of Public Affairs Education, 18(3), 531-544.

Pike, C. K. (1998). A validity study of an instrument designed to measure teaching effectiveness. Journal of Social Work Education, 34, 261–271.

Singer, Judith D. & Willett, John B. (2003). Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence, Oxord University Press.