International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

Vol. 34 No. 2 September 2022, pp. 450-460

Student Level Achievement In Different Public University Of Bangladesh: A Comparison Between Multilevel Models And Classical Regression Models

Anis Mahmud, Mst. Bithi Akter

Department of Statistics, Jahangirnagar University, Savar, Dhaka, Bangladesh

Corresponding Email Address:
anismahmud678@gmail.com
bithiakter200stat@gmail.com

Abstract – In this study, the main purpose is to consider generalized linear modeling where the regression coefficients are also modeled. The models used here are where slopes and intercepts may vary by group or university. The study time of selected students is a dependent variable and the socio and demographic variables are independent variables. The statistical method of data analysis used for this study is the two-level multilevel modeling. In this analysis, we find the dependent variable is associated with time spends in reading academic books (TSRAB), time spends in reading nonacademic books (TSRNAB), and time spends in reading central library (TSRCL) at the 95% level of significance. Using R, we get three multilevel modeling names as the first model is null or empty model when the intercept is fixed, the second model with random intercepts and fixed slopes and the final or the third model with random intercepts and random slopes. Comparing three models, based on AIC and BIC statistics, we get the final model is preferable to the first two models. The results coefficient of determination (R-square) revealed that the multilevel model is better than the classical regression model.

Keywords – Student level, University level, Classical Regression Model, Multilevel Modeling.

I. INTRODUCTION

A classical regression is a relationship between the dependent variable and one or more independent variables. Nowadays, studies are increasingly applying multilevel multivariate research designs. Multilevel analysis is used to examine relations between variables measured at different levels of the multilevel data structure. In multilevel data, also called hierarchically structured or clustered data, lower-order units are clustered within higher-order units, for example, students in a university [4].

Multilevel modeling is a direct way to incorporate indicators for clusters at all levels of a design without being overwhelmed with overfitting problems that arise from applying least squares or maximum likelihood to problems with a large number of parameters [2]. Multilevel data structures may also occur in studies involving repeated measurements collected from abundant individuals [7]. In multilevel data analysis, the need for multilevel modeling is twice. In this study, we used two levels of a multilevel model where one level is students and the other level is university.

Regression models are commonly used for predicting outcomes for new cases. But what if the data vary by group, then we can make predictions for new units in existing groups or in new groups. The latter is difficult to do in classical regression if a model ignores group effects, it will tend to understate the error in predictions for new groups. But a classical regression that includes group effects does not have any automatic way of getting predictions for a new group. For such cases, we can use multilevel modeling. Using multilevel modeling, we will fit a model depending on the study time of different Public University students, departments, faculties and years. Study time (per week)' depends on different levels.

So, the main aim of this research is to analyze the study time of selected students of Public University by using multilevel models along with traditional models and also compare the performance of the competitive models with the multilevel modeling. An overview of regression-type models along with multilevel models with R can be found in [3]. A detailed material and methods related to data source, study variables and statistical models are explained in Section 2. In Section 3, the results and discussion are illustrated. Finally, the conclusion is described in Section 4.

II. MATERIALS AND METHODS

2.1 Source of Data

In this study, the primary data is collected from three different university in Bangladesh named Jahangirnagar University, Mawlana Bhashani Science and Technology University and Comilla University students based on their study time (per week). The survey is conducted on during January 2022 to September 2022. At first, we made a questionnaire to collect the primary data considering the objectives of this study.

2.2 Study Variables

In this study, the study time of selected students is the dependent variable. And the independent variables are gender (male and female), residence (hall and others), name of university Jahangirnagar University (JU), Mawlana Bhashani Science and Technology University (MBSTU) and Comilla University (CU), academic year (category as 1st year, 2nd year, 3rd year, 4th year and masters), CGPA (out of 4), education level of father (category as SSC, HSC, Honors, Masters and others), education level of mother (category as SSC, HSC, Honors, Masters and others), faculty, department, spend time central library (yes or not), time spend in reading academic books (yes or not), time spend in reading nonacademic books (yes or not), expenditure on cell phone, reading newspaper, involvement of cocurricular activities, involvement of extracurricular activities, impact of 1st year result on next year result.

2.3 Multilevel Logistic Regression Model

Consider an educational study with data from students in many universities, predicting in each university's students' grades y on a standardized test given their scores on a pre-test x and other information. The student-level regression and the university - level regression are the two levels of a multilevel model. A model in which the coefficients vary by hall. A model with more than one variance component (student-level and university-level variation), a regression with many predictors, including an indicator variable for each university in the data. A detail about multilevel modeling can be found in [1]. There are two types of multilevel models such as random intercept model and random coefficient model. A details theory of multilevel modeling can be found in [5][6].

The Null Model

The simplest multilevel model has a single residual term for each level. A random intercepts model is a model in which intercepts are allowed to vary, and therefore, the scores on the dependent variable for each individual observation are predicted by the intercept that varies across groups. This model assumes that slopes are fixed. This model is given as:

$$y_{ij} = \gamma_{oo} + \mu_{oj} + e_{0ij}\ ;\ \ e_{0ij} \sim N(0, \sigma_e^2)\ and\ \ \mu_{oj} \sim N(0, \sigma_\mu^2),$$

The intraclass correlation (ICC) is: $\frac{\sigma_\mu^2}{\sigma_e^2 + \sigma_\mu^2}$.

The Final Model

A random slopes model is a model in which slopes are allowed to vary, and therefore, the slopes are different across groups. This model assumes that intercepts are fixed. And a model that includes both random intercepts and random slopes is likely the most realistic type of model, although it is also the most complex. The full model is given by

$$y_{ij} = \gamma_{oo} + \gamma_1 X_{1ij} + \gamma_2 X_{2ij} + \ldots + \gamma_n X_{nij} + \mu_{oj} + e_{ij}\ .$$

To estimate this model parameter, we used maximum likelihood estimation.

III. Results and Discussion

3.1 Exploratory Data Analysis

In our analysis, we are considered n= 200 sample observations and its various characteristics are performed. To fit a multilevel modeling with university wise regression which study time (per week) as dependent variable and other variables are considered as independent variable. Descriptive statistics are useful for describing the basic features of data. In 200 observations, we get some variable, which are associated with dependent variable. The descriptive statistics of our analysis is presented in Table 3.1.

Table 3.1 Descriptive Statistics

Mean	23.42
Standard Error	0.815163
Median	24
Mode	30
Standard Deviation	11.528
Sample Variance	132.8981
Kurtosis	-0.7732
Skewness	0.13652
Range	45
Minimum	2
Maximum	47
Sum	4684
Total	200

The measure of central tendency of three University students with the dependent variable of their study time is measured. Mean is the most widely used measure for central tendency. The average mean value is 23.42 and the median is 24. The mode is 30 which shows a higher frequency in this data. Here, the value of mean and median are almost equal, so we can say that the data are approximately symmetric under the variable study time of the student. In this data the maximum value is 47 and the minimum value is 2, so the range is 45. The standard deviation is 11.528 and the square of the standard deviation is 132.8981 which is called variance. From data, we get the negative value of kurtosis is -0.773 which is less than mean value, so we can say that is a platykurtic. Skewness is measuring the degree of asymmetry of the data and we get the normal skewness which is 0.137.

Table 3.2 Frequency table of the dependent variable

Class of Study time	Frequency	Percentage of the frequency
1-6	5	2.5%
6-12	10	5%
12-18	25	12.5%
18-24	37	18.5%

24-30	56	28%
30-36	40	20%
36-42	16	8%
42-48	11	5.5%
Total	200	100%

A boxplot is a method of graphically depicting groups of numerical data through their quartiles. The boxplot shows the shape, central tendency, and variability of our data.

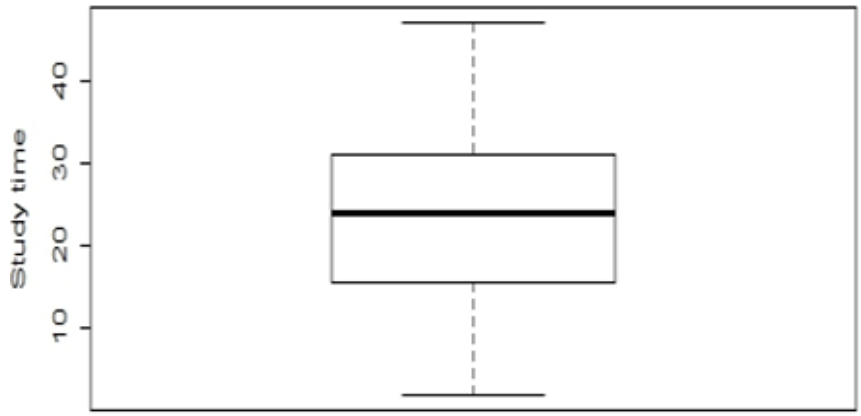

Figure 3. 1 Box and Whisker plot of Study time

Figure 3.1 shows that our data is approximately symmetric and also do not have any outliers.

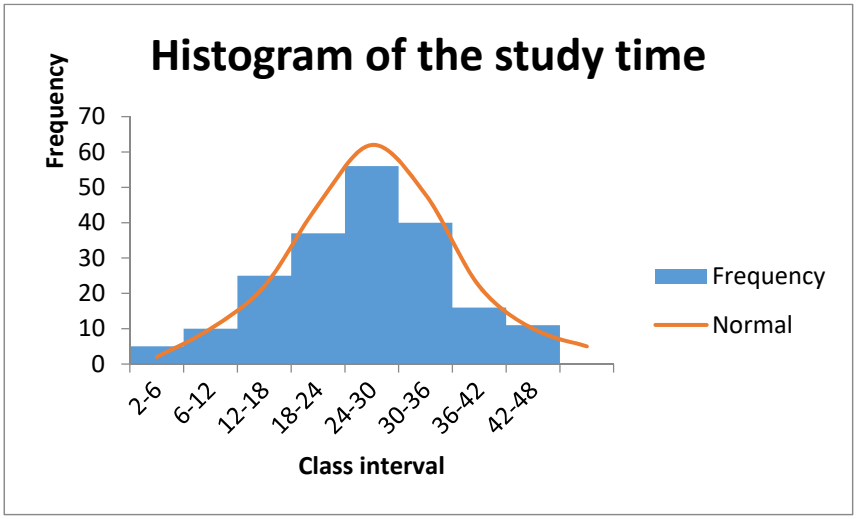

Figure 3.2 Histogram of study time

Figure 3.2 portrayed that the histogram is approximately symmetric and the study time follow the normal curve. A Q-Q plot is easy to interpret and also have the benefit that outliers are easily identified.

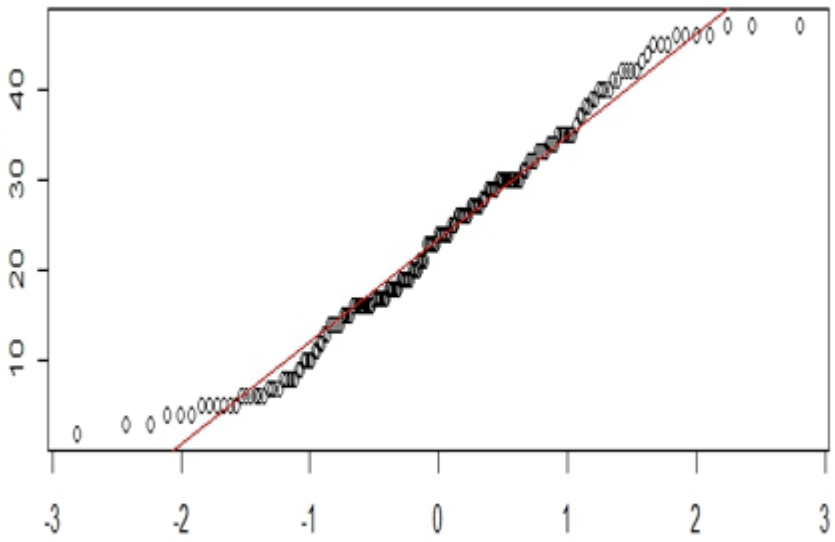

Figure 3.3 Normal Q-Q plot of study time

The Shapiro-wilk normality test, the calculated value is o.9755 and p-value is 0.157. From the figure 3.3, the data appears to be normally distributed. The number associated with each category is called the frequency and the collection of frequencies over all categories gives the frequency distribution of that variable**.** Table 3.3 summarizes the distribution of values in the sample and also perform the frequency distribution in different characteristics of the student, depends on the study time (per week). In 200 students the 52 percent is male and 48 percent is female. Above the 200 student's 99.5 percent stay in hall but 0.5 percent is not, total students selected from JU, MBSTU and CU are 97, 38, and 65 respectively and similarly interpreted others variables.

Table 3.3 Frequency Table with major variable

Variable	**Category**	**Frequency**	**Percentage**
Gender	Male	104	52.0
	Female	96	48.0
Residence	Hall	199	99.5
	Others	1	.5
Name of university	JU	97	48.5
	MBSTU	38	19
	CU	65	32.5
Education of father	SSC	39	19.5
	HSC	41	20.5
	Honors	54	27
	Masters	45	22.5
	Others	21	10.5

Education of mother	SSC	85	42.5
	HSC	31	15.5
	Honors	21	10.5
	Masters	14	7
	Others	49	24.5
Year	Second year	18	9
	Third year	30	15
	Fourth year	44	22
	Masters	108	54
Impact of 1st year result on your next year result	Yes	161	80.5
	No	39	19.5
Tendency of going central library	Yes	85	42.5
	No	115	57.5
Having reading room in hall	Yes	101	50.5
	No	99	49.5
Reading newspaper	Yes	91	45.5
	No	109	54.5

Table 3.4: Summary of cross tabulation between the variables that are statistically significant at 95% confidence interval.

Dependent variable	Independent variable	Pearson chi-square	p-value	Decisions (at 5% level of significance)
Study time (per week)	Family member	326.301	0.227	Not associated
	Year	161.293	0.04	Associated
	CGPA	29411.121	.142	Not associated
	Education of father	191.194	0.203	Not associated
	Faculty	231.626	0.925	Not associated
	Department	1416.983	0.424	Not associated
	Spend Time central library	1476.093	0.018	Associated
	Time spends in reading	1436.670	0.000	Associated

	academic books			
	Time spends in reading nonacademic books	2588.09	0.000	Associated
	Expenditure on cell phone	1273.505	.303	Not associated
	Education of father	1241.09	.527	Not associated
	Education of mother	192.02	.303	Not associated
	Involvement of cocurricular activities	197.622	.126	Not associated
	Involvement of extracurricular activities	169.182	.63	Not associated

By performing cross-tabulation we have to find out the relationship between study time (dependent variable) and other independent variables from our total observations and also see that they are related or not by using chi-square distribution. From Table 4.4, the variables are mentioned as statistically significant at 5% level of significance. Others variables are found insignificant at 5% level of significance, using the spreadsheet software SPSS. The statistic of p-value means if the p value of all variables is less than 0.05, we may reject the null hypothesis. That means there is a relationship between two variables. So, we may conclude that the dependent variable which is study time is associated with time spends in reading academic books (TSRAB), time spends in reading nonacademic books (TSRNAB), and time spends in reading central library (TSRCL).

Now, we perform classical regression model to explore only the conditional relationship between dependent variable and three independent variables. The estimated models are given bellow and the variables are denoted by: Y= study time (per week); X_1= time spends in reading academic books (TSRAB); X_2= time spends in reading nonacademic books (TSRNAB); and X_3= time spends in reading central library (TSRCL). So, the linear regression model is, for i= 1, 2, …, n

$$\hat{y} = 0.165552 + 0.992807\, x1 + .98880\, x2 + 0.003292\, x3,$$

where, R^2 = 0.87, RSS = 32.6, df = 196 and n = 200.

3.2 Estimation of Multilevel Modeling

3.2.1 Estimation with R

In this section we have to discuss how to center variables and estimate multilevel models using R. For this estimate, we have to install and load the package lme4, which fits linear and generalized linear mixed-effects models. By using R, we get the following result which is given following table.

Table 3.5: Comparison among three model by multilevel Modeling (result from R)

Fixed Effects	Model 1	Model 2	Model 3
Intercept (γ_{oo})	23.91	12.70207	0.391540
Time spends in reading academic books (γ_{10})			0.987151
Time spends in reading nonacademic books (γ_{20})			0.982464
Time spends in reading central library (γ_{01})		0.58777	0.003667
Random Effects	Model 1	Model 2	Model 3
Intercept (μ_{oj})	41.56 (6.447)	0.00 (0.000)	0.2907757 (0.53924)
Time spends in reading academic books (μ_{1j})			0.0003106 (0.01762)
Time spends in reading academic books (μ_{2j})			0.0002131 (0.01460)
Residual (e_{ij})	88.18 (9.391)	45.18 (6.722)	0.1476698 (0.38428)

Model Fit Statistics	Model 1	Model 2	Model 3
Deviance	1473.7	1329.7	194.3
AIC	1479.7	1337.7	216.3
BIC	1489.6	1350.9	252.5

Null Model

The function of the model in hall wise object which is estimated our model with study time is the dependent variable. In given data, a model for all effects on student's study time there would be student (level-1), university (level-2). Empty or null model is the intercept for the fixed effects, the random effects intercept and the variable of university is specified as the level-2 grouping variable.

$$\hat{y}_{ij} = \beta_{oj} + e_{ij}$$

$$\beta_{oj} = \gamma_{oo} + \mu_{oj} \quad = \quad 23.91 + 41.56$$

$$\hat{y}_{ij} = \gamma_{oo} + \mu_{oj} + e_{ij} = \quad 23.91 + (41.56 + 88.18)$$

From this model, the average test score across universities, reflected in the fixed effects intercept term, is 23.91. The variance component corresponding to the random intercept is 41.56. The intraclass correlation coefficient is equal to $\frac{41.56}{88.18+41.56}$ =0.32, meaning that roughly 32% of the variance is attributable to the university -level.

Models with random intercepts and fixed slopes

With the introduction of the student-level variable time spend reading central library as a fixed effect, the equation can be written as,

$$\hat{y}_{ij} = \beta_{0j} + \beta_{1j} \text{ (Time spend in reading central library)} + e_{ij}$$

$$\beta_{oj} = \gamma_{oo} + \mu_{oj} = 12.70207 + 0.00 = 12.70207$$

$$\beta_{1j} = \gamma_{01} \text{ (TSRCL)} = 0.58777 \text{ (TSRCL)}$$

$$\hat{y}_{ij} = \gamma_{oo} + \gamma_{01}(\text{TSRCL}) + \mu_{oj} + e_{ij}$$

$$= 12.70207 + 0.58777 \text{ (TSRCL)} + 45.18$$

The intercept, which now corresponds study time in any one of the universities with average reading score is 12.70207. The fixed term of TSRCL score is 0.58777. The variance component corresponding to the random intercept has decreased to 0.00, that inclusion of the level-2 variables has accounted for some of unexplained variance in the study time (per week). The estimate still more than twice the size of its standard error, suggesting that there remains unexplained variance.

Models with random intercepts and random slopes

In the cases examined so far, the within university regression lines were all parallel, but multilevel regression analyses also allowed the regression slopes to vary. The effect is that the reading academic book and nonacademic book effect, will be considered as random, while in the latter, the effect will be considered as fixed.

$$\hat{y}_{ij} = \beta_{0j} + \beta_{1j} \text{ TSRAB} + \beta_{2j} \text{ TSRNAB} + e_{ij}$$

$$\beta_{oj} = \gamma_{oo} + \gamma_{01} TSRCL + \mu_{oj} \quad = \quad 0.391540 + 0.003667 \text{ (TSRCL)} + 0.2907757$$

$$\beta_{1j} = \gamma_{10} TSRAB + \mu_{1j} TSRAB \quad = \quad 0.987151 \text{ (TSRAB)} + 0.0003106 \text{ (TSRAB)}$$

$$\beta_{2j} = \gamma_{20} TSRNAB + \mu_{2j} TSRNAB \quad = \quad 0.982464 \text{ (TSRNAB)} + 0.0002131 \text{ (TSRNAB)}$$

Full Model,

$$\hat{y}_{ij} = \gamma_{oo} + \gamma_{o1} \text{ (TSRCL)} + \gamma_{10} \text{ (TSRAB)} + \gamma_{20} \text{ (TSRNAB)} + \mu_{1j} (TSRAB) + \mu_{2j} \text{ (TSRNAB)} + \mu_{oj} + e_{ij}$$

$$= 0.391540 + 0.003667 \text{ (TSRCL)} + 0.987151 (TSRAB) + 0.982464 (TSRNAB) + 0.0003106 (TSRAB) + 0.0002131 (TSRNAB) + (0.2907757 + 0.1476698)$$

From the above, the variance component for the random intercept is 0.2907757, which is still large relative to its standard deviation of 0.53924. Thus, some university-level variance remains unexplained in the final model. The variance component corresponding to the slope, however, is quite small relative to its standard deviation. This suggests that we may be justified in constraining the effect to be fixed. Table 3.5 displays the deviance and AIC and BIC. Comparing both the AIC and BIC statistics, it is clear that the final model is preferable to the first two models.

3.2.2 Coefficient of determination (R^2)

The proportion of response variable variance accounted for by the model is expressed as R^2. In the context of multilevel modeling, R^2 values can be estimated for each level of the model.

For level-1, we can calculate

$$R^2 = 1 - \frac{\sigma^2_{u1} + \mu^2_{u1}}{\sigma^2_{u0} + \mu^2_{u0}}$$

$$= 1 - \frac{0.38428 + 0.53924}{9.391 + 6.447}$$

$$= 0.94169$$

This result tells us that Level-1 of specified model explains approximately 94% of the variance in the reading academic and nonacademic book above and beyond that accounted for in the null model.

We can also calculate a level-2 R^2 value:

$$R^2 = 1 - \frac{\sigma^2_{u1/B} + \mu^2_{u1}}{\sigma^2_{u0/B} + \mu^2_{u0}}$$

$$= 1 - \frac{0.545}{6.588}$$

$$= 0.917$$

Where, B is the average size of the level 2 units. R provides the number of individuals in the sample (n=200) and the number of university (3). So that we can calculate B as 200/3= 66.67. The level-2 of specified model explains approximately 92% of the variance in the null model.

Multilevel modeling is an increasingly popular approach to modeling hierarchical structured data, outperforming classical regression in predictive accuracy. In regression analysis, the variability of response variable (R^2) is smaller than the variability of response variable (R^2) of multilevel modeling.

IV. Conclusion

In classical linear regression models, the parameters of the model are constant and explanatory variables are fixed but, in some situations, where the response variable depends on different level of the explanatory variables or by itself. In that situation, the regression coefficients can vary as well as intercepts. In such situation multilevel model is more appropriate.

In this study, we analyzed the study time of the students with three University in Bangladesh with its related factors via regression type model. As we mentioned, we used 2-level of multilevel model. The dependent variable of study time is associated on time spends in reading academic books, time spends in reading nonacademic books, and time spends in central library. We observe that study time can vary among the university student. Thus, we consider, multilevel model instead of classical linear regression model to analyze study time.

We wish to test whether the slope is random on three university students where the slopes are spent time in reading academic book, spend time in reading nonacademic book, and spend time in central library. Table 3.5 shows that the model 3 is better than the others two models based on the minimum value of AIC, BIC and deviance.

Finally, we calculate that the coefficient of determination (R^2) for multilevel model and classical linear regression model and we obtain that the coefficient of determination (R^2) for multilevel model is 92 percent and the coefficient of determination (R^2) for classical linear regression model is 87 percent. So, it is clear that multilevel model is better than classical linear regression model.

Authors Contribution

The research was conducted by Anis Mahmud. Mst. Bithi Akter and Anis Mahmud collected and implementation the data. Anis Mahmud analyzed the data and shared the SPSS and R code with Mst. Bithi Akter. Anis Mahmud finalized the data analysis and tabulated the results and wrote a draft copy of the manuscript. Mst. Bithi Akter finalized the manuscript.

Funding

The author(s) received no financial support for the study, authorship, and publication of this article.

Data and Code Availability

The data sources are provided the Section 2. We will provide the data and R-code if anyone needs it.

Ethics Approval

We have worked with honesty and devotion. We have not knowingly engaged in or participated in any harm to another person or animal.

Conflict of Interest

The authors declare that they have no conflicts of interest

REFERENCES

[1] Hox J, Moerbeek M, and Van de Schoot R (2018). Multilevel Analysis, Techniques and Applications, Third Edition.

[2] Gelman, A. and Hill, J. (2006). Applied Regression and Multilevel/Hierarchical Models. Cambridge University Press.

[3] Finch, W.H., Bolin, J.E., and Kelley, K. (2014). Multilevel modeling using R.

[4] David, M. M.., and Fesperman, L. (2013). Advanced Standard SQL Dynamic Structured Data Modeling and Hierarchical Processing. Artech House, Boston.

[5] O'Connell, Ann A and McCoach, D. Betsy (2008). Multilevel modeling of educational data. Charlotte, NC, IAP Publishers.

[6] Snijders, T., and Bosker, R. (2012). Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. London: Sage Publishers.

[7] Ryu, E. (2014). Model Fit Evaluation in Multilevel Structural Equation Models. Frontiers in Psychology, volume 5.