A MULTILEVEL LONGITUDINAL ANALYSIS OF TEACHING EFFECTIVENESS IN CASE OF MONGOLIAN PRIVATE UNIVERSITY

Uilst Dorligsuren, PhD candidate, Beijing Normal University
Myagmarsuren Boldbaatar, PhD, associate professor, National University of Mongolia

Хураангуй

Энэ судалгаа нь Монголын нэгэн хувийн их сургуулийн тасралтгүй 11 жилийн хугацаанд цуглуулсан оюутны үнэлгээний мэдээлэл дээр тулгуурлан сургалтын үр ашгийн өсөлтийг судалсан. Оюутны үнэлгээний дата нь 2005-2006 оны хичээлийн жилээс 2015-2016 оны хичээлийн жилийн хугацаанд авсан нийт 62 багшийг үнэлсэн 87,555 оюутны үнэлгээний асуулгыг багтаасан бөгөөд олон шатлалт урт хугацааны шинжилгээний аргыг ашиглан дүн шинжилгээ хийсэн болно.

Судалгааны үр дүнгээс харахад 11 жилийн хугацаанд тус их сургуулийн сургалтын үр өгөөж нь статистик ач холбогдолтойгоор сайжирсан байна. Мөн багшилсан жил, багшийн нас, хүйс, боловсролын түвшин зэрэг нь оюутны үнэлгээнд нөлөө үзүүлдэггүй болох нь тодорхойлогдсон. Сургалтын үр өгөөж нь цаг хугацааны явцад өөрчлөгдөж, багш нарын сургалтын үр өгөөж хоорондоо харилцан адилгүй байна. Судалгааны эхэн үед оюутны үнэлгээгээр доогуур үнэлгээтэй байсан багш нар илүү өндөр үнэлгээтэй байсан багш нартай харьцуулахад илүү их өсөлттэй байна. Багшийн сургалтын ажлыг үнэлэх оюутны үнэлгээнд хичээлийн төрөл, оюутны тоо, хичээлийн улирал зэрэг хүчин зүйлүүд нь мөн чухал нөлөө үзүүлдэг байна.

Түлхүүр үгс: сургалтын үр ашиг, багшийн үнэлгээ, оюутны үнэлгээ, олон шатлалт урт хугацааны шинжилгээ

Abstract

This study examined the growth trajectory of teaching effectiveness at one particular higher education institution in Mongolia based on student rating data collected over 11 consecutive years. The student rating data covers 87,555 questionnaires evaluating 62 teachers rating collected over the period from 2005-2006 academic year to 2015-2016 was analyzed employing a multilevel longitudinal analysis method.

The results show that that the teaching effectiveness of the private university in Mongolia which was selected in this study improved significantly over an 11-year period. Teaching experience, age, gender and academic degree which are three measures of seniority found to have no significant effect on the student ratings. The study findings also reveal that teaching effectiveness varies over time and that teachers differ from each other in terms of teaching effectiveness. Moreover, there is a strong negative relationship between the initial status and the rates of change which means teachers who have lower starting points tend to have higher growth rate compared to teachers with higher starting points. The three variables of course type, number of students, and semester have the significant relationship with the student ratings of teaching effectiveness.

Key words: teaching effectiveness, faculty evaluation, student ratings, multilevel longitudinal analysis

1. INTRODUCTION

The quality of Mongolian higher education has been a major concerning issue in the last two decades. As evaluation is a critical component of quality assurance in higher education, it is important for any higher education institution to develop an effective faculty evaluation system.

Student ratings of instruction (SRI) are widely used by higher education institutions throughout the world as part of their faculty evaluation system. Student ratings are considered as the most reliable and valid source of data collected on instructor's teaching effectiveness among multiple sources of information which are used in faculty evaluation (Benton and Cashin, 2014; McKeachie, 1979). Therefore, higher education institutions in Mongolia have modified their faculty evaluation programs in the recent years and adopted similar faculty evaluation system used in western countries to assess faculty and teaching effectiveness based on student ratings.

Although this type of teaching evaluation method has been widely used in western countries for almost a century now, Mongolian higher education institutions have recognized the significance of student opinions in teaching

evaluation and started using them only in the recent two decades. Since student ratings have been employed in Mongolian higher education for a relatively short period, studies and resources on this topic are quite rare in Mongolian context.

This study is significant because it is the first longitudinal study that focuses on investigating teaching effectiveness based on students' feedback in the context of Mongolia. There is a lack of published work relating to higher education in Mongolia particularly studies on students' evaluation of teaching effectiveness are almost non-existent. Therefore, this study will make a significant contribution to the limited literature on Mongolian higher education, faculty, and faculty evaluation studies which will be valuable to many stakeholders including academics and international and comparative education researchers. The study results would allow the researcher to test whether results for a Mongolian case study are consistent with consolidated findings from other previous longitudinal studies on student ratings of teaching effectiveness.

In this study we attempted to study the growth trajectory of teaching effectiveness in one particular higher education institution in Mongolia based on the student ratings data collected over 11 consecutive years employing a multilevel longitudinal analysis method. Therefore, the analysis aims to investigate whether the teaching effectiveness at the selected university has improved from 2005 to 2016 and examine the growth trend if there is an improvement as well as if there are individual differences among faculty members in terms of the growth trajectory of teaching effectiveness.

It examines whether the student ratings of teaching performance are influenced by certain variables related to teacher characteristics including gender, age, academic degree and teaching experience of the faculty, and school which the faculty belongs as well as course characteristics such as course type (major v s. general), class size and the semester which the course was provided.

2. LITERATURE REVIEW

The literature review begins with a brief study of the international practices of faculty evaluation followed by the student ratings as a part of faculty evaluation along with its reliability and lastly current practices of evaluation of faculty in Mongolian higher education.

International practices

In higher education institutions, the quality of education is dependent on the faculty performance. Therefore, faculty evaluation plays an important role in quality assurance in higher education "because teacher evaluation is at the heart of the educational enterprise-the quality of teaching in the nation's classrooms-it has the potential to be a powerful lever of teacher and school improvement" (Toch & Rothman, 2008).

Faculty evaluation is generally based on performance in three main areas: teaching, research, and service (Arreola, 2000). Evaluating teaching performance has different forms including student evaluation, peer evaluation, self-evaluation and evaluation by academic administrators. Many institutions use a combination of these methods to evaluate their faculty's teaching performance (Seldin, 1984). Based on the nationwide surveys among 616 accredited, four-year, undergraduate, liberal arts colleges to examine the policies and practices in the evaluation of faculty performance, Seldin (1984) reported that approximately 99 percent of the academic deans who participated in the survey consider teaching as the most important index of evaluating overall faculty performance.

As a result, SRIs have become a routine procedure in higher education institutions worldwide and an important element in the faculty evaluation process. Student ratings of instruction was introduced into North American universities in the mid-1920s (d'Apollonia & Abrami, 1997). Feldman (2007) noted that the use of student ratings would continue to grow in the future due to the increased emphasis institutions are putting on effective teaching.

As Hoyt and Pallett (1999) highlihgt that most universities and colleges employ student ratings as part of the evaluation of teaching performance due to relatively simple procedure of collecting student feedback and credibility. Considerable evidence indicates that students, if asked the right questions that relate to their frames of reference, are valid and reliable judges of teaching effectiveness. Seldin (1984) points out students are an excellent source in terms of assessing instructor's instructional skills. Murray (1997) also supported the use of student ratings forms to assess teacher and course

characteristics such as clarity of explanation, enthusiasm for subject matter, encouragement of student participation, breadth of coverage, and quality of feedback as student ratings are assumed to be "observable by students; under the control of the instructor; and correlated with student learning." (p. 8).

Researchers suggested that SRIs have the following main purposes: (1) providing faculty with formative[20] feedback about their effectiveness for improving teaching, course content and structure; (2) providing administrators with summative feedback[21] for personnel decisions; (3) providing students with information for the selection of courses and teachers, (4) providing researchers with information for research purposes, (5) providing evidence for institutional accountability (Marsh & Dunkin, 1997; Hativa 2014 (a).

Validity and reliability are two important concepts which have been the focus of hundreds of studies on SRIs. Validity is the extent to which an instrument measure what it is designed to measure. Reliability is the extent to which an instrument is consistent in measuring whatever it is measuring. In the case of student evaluation of teaching effectiveness, the questions related to validity are: 'Do student ratings measure teaching effectiveness?', 'To what extent do they measure the aspects of teaching effectiveness?', and 'Are student ratings biased?' On the other hand, reliability studies generally address the question 'Are student ratings consistent over time and among different raters?'

Thousands of studies have been carried out on the validity and reliability of SRIs. For the most part, the literature supports the reliability and validity of student ratings. Murray (1997) reported that over 1,500 published studies indicate student ratings can provide reliable and valid evidence of teaching effectiveness. Cashin (1995) notes "In general, student ratings tend to be statistically reliable, valid, and relatively free from bias or the need for control; probably more so than any other data used for faculty evaluation" (p.6). On the basis of a review of literature, Seldin (1984) concluded that "In the matter of student ratings, virtually every study measuring their reliability has reported a high level of stability and consistency." (p.134) Based on a number of extensive studies of validity and reliability of SRIs, Marsh and Roche (1997) concluded student ratings are (a) multidimensional; (b) reliable and stable; (c) primarily a function of the instructor who teaches a course rather than the course that is taught; (d) relatively valid against a variety of indicators of effective teaching; (e) relatively unaffected by a variety of variables (f) useful in improving teaching effectiveness when they are used with appropriate consultation.

The reliability of SRIs is most appropriately determined from studies of inter rater agreement that assess agreement among different students within the same course. Overall and Marsh (1980) conducted a longitudinal study to examine the agreement between responses by current and former students. In their findings, ratings in 100 classes correlated .83 with ratings by the same students when they again retrospectively evaluated the same classes several years later. When examining the effects of the teacher and the course on the SRIs, Marsh (1987) found out that the correlation between overall ratings of different instructors teaching the same course whereas correlations for the same instructor in different courses and in two different offerings of the same course were much larger. These results provide support for the validity of SRIs as a measure of teacher effectiveness not the course effectiveness.

Brief introduction of faculty Evaluation in Mongolian Higher Education

As of 2016, there are 100 higher educational institutions including 17 public, 78 private and 5 foreign universities and colleges. A total of 162,626 students are enrolled and 7,121 faculty members are employed (Ministry of Education, Culture and Science, 2016). Due to the rapid expansion, there is an increasing competition among the higher education providers. Both public and private universities and colleges are now considering much more carefully how they can achieve a competitive advantage to attract more students, and this has led many institutions to take an increased interest in student satisfaction.

[20] In formative evaluation, student ratings are used to provide beneficial feedback to the teachers and assist them make improvements in their teaching performance (McKeachie, 1997).

[21] In summative evaluation, student ratings are used to inform administrative decisions regarding the merit, worth, or value of the instructor's teaching abilities (Theall & Franklin, 2001).

In the recent years, Mongolian universities and colleges have modified their faculty evaluation regulations. Faculty evaluation is conducted in similar ways in western universities and colleges. In most Mongolian universities and colleges, faculty evaluation is based on three main areas: teaching, research and service (professional and social). Student evaluation of teaching was introduced in the evaluation system in the late 1990s at some universities and used in most higher education institutions.

Mongolian public and private higher education institutions have established their own faculty evaluation system. For example, the faculty evaluation regulations of the National University of Mongolia[22] (2012) notes that faculty members will be evaluated based on three main areas: teaching, research and service (professional and social). Depending on the academic rank, different weights are given to the faculty member's contributions in teaching, research, and service. When evaluating teaching and pedagogy, student evaluation takes up 40 percent of the evaluation.

Although we have rich historical data on the student evaluation, almost none analysis has been made on these data in our country.

Internationally, literature of SRIs is abundant as this has been one of the most studied topics in higher education in the last several decades, most of the studies are cross-sectional and less number of longitudinal studies on this topic are currently available. Particularly, there are just a few longitudinal studies which used the multilevel model which ideally suited to this type of data. Given that there is a limited number of longitudinal studies in this area as well as the mixed results of the below studies, there is a definite need for more longitudinal studies in this field, especially applying the multilevel model.

Murray and others (1996) conducted a large-scale longitudinal study which analyzed the student ratings of 40 to 50 full-time faculty members in the Department of Psychology at the University of Western Ontario collected over 26 consecutive years between 1970 to 1995. The study found a significant improvement in teaching effectiveness over a period of 26 years.

Lang and Kersting (2006) examined whether feedback from student ratings of instruction not supplemented with consultation helps teachers to improve their ratings on a long-term basis. A sample of 3122 questionnaires evaluating 12 teachers from the psychology department at a large German university collected over a 4-semester period were analyzed in the study. The results revealed that student ratings increased from the no-feedback baseline semester to the second semester and then gradually declined from the second to the fourth semester. The researchers concluded that consultation interventions could affect long-term improvements in teaching effectiveness.

Marsh (2007) conducted a longitudinal study examining the stability of university teaching effectiveness by applying multiple-level growth modeling approach on the basis of SRIs for a cohort of teachers who were evaluated continuously over a 13-year period from 1976 to 1988 to investigate whether teaching effectiveness increases, decreases, or remains stable with added experience. The study involved a diverse cohort of 195 teachers from 31 departments at the University of Southern California. This study showed that there was little evidence that teachers became either more or less effective with added experience. Marsh concluded that without systematic intervention, teaching effectiveness at all levels regardless of different methods of measurement tends to decline with age and years of teaching experience.

By using the same method, Carle (2009) analyzed the data from 10,392 classes taught by 1120 teachers across three years to examine whether students' ratings of teaching effectiveness changed across time, whether differences in average student ratings correlated with growth, and whether certain course and teacher characteristics affected the ratings. The study found that the students' ratings of teaching effectiveness remained relatively stable across time and teachers, although analyses revealed a statistically significant, negative correlation between initial status and growth which instructors starting with lower ratings improved the fastest. Factors such as discipline, course level, gender, minority status and tenure have no significant effects on teaching effectiveness.

[22] Pioneer of higher education in Mongolia

Hallinger (2010) conducted a longitudinal case study research at a university in Thailand which analyzed the student course evaluation data of 233 teachers evaluated by 40,686 students at a graduate school of business gathered over a period of 21 terms during a seven-year period. The study found statistically significant improvement in levels of teaching effectiveness.

Bianchini and others (2012) examined the longitudinal dataset from an Italian University to investigate the relationship between students' evaluation of teaching effectiveness and teacher's characteristics including age, gender, academic rank, and disciplinary affiliation using multivariate regression analysis. Student ratings data of three years was used in the study. Age and academic rank were both found to affect negatively teaching evaluation. Profession-oriented disciplines were evaluated lower. Gender was also found to be relevant as female teachers were rated lower than their male counterparts. Past research activities as measured by the number of publications of instructors, has a positive impact on student ratings. As for rank and academic discipline, full professors and external faculty got lower ratings.

3. METHODOLOGY AND DATA

Most previous studies (international) on student ratings have considered ratings of teaching performance collected in one specific course on a single occasion and there is a limited research on the stability of student ratings over an extended period of time. However, we employ a multilevel longitudinal analysis method in this study. Because cross-sectional studies do not provide a strong basis for understanding how ratings of the same instructor vary over time and determining the long-term effects of certain instructor variables on teaching effectiveness (Marsh, 2007).

As Murray (1997) suggested certain methodological conditions need to be fulfilled when conducting a longitudinal study. Those are as follows:

- mean ratings are compared across a minimum of ten years or ten semesters;
- tracking of mean ratings across years begins in the same year that student evaluation was first introduced;
- the same student rating form is used throughout the study; and
- all faculty and all courses undergo student evaluation in all years.

The data obtained from the selected university which was used in the study fulfills all of the methodological conditions identified above. The archival data used in this study are student ratings collected over 11 consecutive years or 22 semesters. The student ratings were first introduced in the selected university in 2005 and the data covered from 2005 to 2016 is used in this analysis. The same seven-item student rating form has been used continuously throughout 11-year period. All faculty members are evaluated by their students at the end of every semester.

3.1 Data Source

The university which was selected to conduct the study was established in 1991 Ulaanbaatar Mongolia. It is one of the pioneer private universities of the country. The university has expanded over the years and currently it has three branch schools: School of Humanities, School of Law and School of Business offering undergraduate (BA) and graduate (MA, PhD) degrees in a variety of fields. It currently enrolls over 2,000 students and employs about 90 faculty members.

Data used in this study were obtained from the Academic Affairs Office and the Archival Office at the selected private university in Mongolia. The student ratings data of each semester have the following information: (1) instructor's name, (2) number of students participated in the survey, (3) averaged rating score for each indicator of instruction being evaluated, and (4) mean scores of the ratings of the seven indicators.

Additional information related to teacher characteristics including age, gender, years of teaching experience, academic degree, and school that the instructor belongs, was obtained from the Office of the Academic Affairs Office. For anonymity reasons, all teachers used in the study were coded and their names were removed from the data in the initial stage of data processing.

The university's student ratings data kept in the archive show that a total of 209 teachers worked during 2005-2016. Originally, it was intended to include all teachers who worked during this time frame in the study. As the study aims to investigate the difference in the ratings of

teaching performance related to certain teacher and course characteristics, it is preferable to have as larger sample size as possible. Since this study is a longitudinal study, it is not appropriate to include teachers who worked only a few semesters as some of the teachers worked only one or two semesters. Therefore, the teachers who worked less than ten semesters were eliminated from the study.

The sample considered in this survey consists of all teachers who were evaluated in ten or more semesters over an 11-year period. The resulting group included 62[23] full-time faculty members who have worked more than five years or ten semesters at this institution and who come from seven different academic departments of the selected university. Instructors at this university teach 1-3 different courses each semester. Each instructor's rating scores for different courses are aggregated in each semester; therefore, each teacher has one rating score for one semester regardless of the number of courses he or she taught.

In order to test the internal consistency reliability of the dataset we calculate Cronbach's alpha. It was 0.954 which indicates that it is acceptable for our research purposes.

3.2 Data description

The dependent variable or the outcome variable in all analyses was the class-average rating or the overall teacher rating averaged across all responding students in all classes. The predictor variables include time, gender, age, academic degree, teaching experience (years of teaching experience at the first measurement), course type, number of students evaluating each teacher every semester, and semester which the course was provided. Detailed descriptions are in Table1.

The dependent variable is the class-average rating – the overall teacher rating averaged across all responding students who evaluated the same teacher in each semester. One considerable factor in the data was that the number of students evaluating teachers each semester varied substantially, ranging from 4 to 553; 1.5% of the class-average responses were based on fewer than 10 students.

Marsh (1987) indicated that the number of students responding within each class is a factor that should be taken into consideration in the SRIs research, particularly in the evaluation of stability. He suggested that the reliability of class-average ratings varies with the number of students per class; 0.95 for 50 students, 0.90 for 25 students, 0.74 for 10 students, and 0.60 for 5 students. In our study, 98.5 percent of the class-average responses were based on more than 10 students. Hence the reliability of class-average ratings is relatively high. The descriptive statistics for the independent variables were provided below.

Independent variables related to faculty characteristics (gender, age, academic degree, teaching experience, and school in which the faculty belongs) and course characteristics (course type, the number of students who evaluated the instructors, and the semester of the student ratings) are presented in Table 1.

[23] 87,555 questionnaires evaluating the 62 teachers in 11 consecutive years or 22 semesters.

Table 1. Descriptive statistics of the variables

Variable	Obs.	Mean	Std. Dev.	Min	Max	Description
Overall teacher rating	991	90.05	5.32	64.6	99.4	Dependent variable
Gender	991	0.63	0.48	0	1	0-male; 1- female
Age	991	36	10.30	20	76	in whole years
Academic degree	991	0.17	0.38	0	1	0- master degree; 1- doctoral degree
Teaching experience	991	10.46	8.77	1	52	Years of teaching experience at the first measurement
Course type	991	0.32	0.47	0	1	0-major course, 1- general course
Number of Students	991	88.35	60.40	4	553	Number of students who evaluated one particular teacher
Semester	1364	0.50	0.50	0	1	0- fall semester, 1-spring semester
School 1	991	0.46	0.50	0	1	Dummy variable: 1 – School of Humanities, 0- other
School 2	991	0.23	0.42	0	1	Dummy variable:1 – Law School, 0- other Baseline category indicates School of Business
Time	1364	10.50	6.50	0	21	22 semesters a coded between 0 and 21.
Early	991	0.07	0.25	0	1	Dummy for Early career status: 0- worked until 3 years, 1- worked more than 3 years.

3.3 Research Questions

This study attempts to answer next questions:

- Did the teaching effectiveness of faculty members at this university improve over time?
- Do faculty members at this university share similar growth trend in student ratings of teaching effectiveness?
- Are ratings of teaching performance influenced by certain variables related to teacher characteristics including gender, age, academic degree, teaching experience, and school that the teacher belongs, as well as course characteristics including course type, class size and the semester which the course was provided?

3.4 Model specification

The equations of the two levels of analysis for our data analyses would be:

Level-1 submodel (repeated measures):

$$y_{ti} = \pi_{oi} + \pi_{1i} Time_{ti} + e_{ti}$$

Level-2 submodel (subject level):

$$\pi_{0i} = \beta_{00} + \beta_{01}\text{Gender}_i + \beta_{02}\text{Age}_{ti} + \beta_{03}\text{Academicdegree}_i + \beta_{04}\text{Teachingexperience}_i + \beta_{05} Coursetype_i + \beta_{06} Numberofstudents_{ti} + \beta_{07} Semester_{ti} + \beta_{08} School_i + \beta_{09} Early_i + \gamma_{00i}$$

$$\pi_{1i} = \beta_{10} + \beta_{11}\text{Gender}_i + \beta_{12}\text{Age}_{ti} + \beta_{13}\text{Academicdegree}_i + \beta_{14}\text{Teachingexperience}_i + \beta_{15} Coursetype_i + \beta_{16} Numberofstudents_{ti} + \beta_{17} Semester_{ti} + \beta_{18} School_i + \beta_{19} Early_i + \gamma_{10i}$$

Below is the single equation for this research which combined the two levels:

$$y_{ti} = \beta_{00} + \beta_{01}\text{Gender}_i + \beta_{02}\text{Age}_{ti} + \beta_{03}\text{Academicdegree}_i + \beta_{04}\text{Teachingexperience}_i + \beta_{05} Coursetype_i + \beta_{06} Numberofstudents_{ti} + \beta_{07} Semester_{ti} + \beta_{08} School_i + \gamma_{00i} + \beta_{10} Time_{ti} + \beta_{11}\text{Gender}_i Time_{ti} + \beta_{12}\text{Age}_{ti} Time_{ti} + \beta_{13}\text{Academicdegree}_i Time_{ti} + \beta_{14}\text{Teachingexperience}_i Time_{ti} + \beta_{15} Coursetype_i Time_{ti} + \beta_{16} Numberofstudents_{ti} Time_{ti} + \beta_{17} Semester_{ti} Time_{ti} + \beta_{18} School_i Time_{ti} + \gamma_{10i} Time_{ti} + \beta_{19} Early_i Time_{ti} + e_{ti}$$

Where:

- is time and is the individual sample in the data.
- represents the overall teacher rating for individual at the time of
- is the intercept of the regression which represents individual *i*'s true initial status, the value of the outcome when.
- is the slope which represents individual *i*'s true rate of change during the period under study.
- represents that portion of individual *i*'s outcome that is unpredicted at the time of , in other words it is the unestimated residual indicating the variability of the data around the regression line.
- the level-2 intercepts, represent the population average initial status and rate of change.
- represents the variance between individual intercept and the average intercept
- represents the variability of individual rate of change around the average population rate of change.

Statistical analyses are conducted using STATA and SPSS packages. A total of 12 multilevel models were fitted, allowing the amount of variance explained at each level to be calculated with the addition of more variables.

4. ESTIMATION RESULTS

The results from a set of two-level longitudinal growth models are based on several combinations of predictor variables. Each of the fit models include two levels: Level 2 = teacher, Level 1 = time. All results are aggregated in Table 2 at the appendix.

Model 1: Unconditional Means Model

When applying multilevel models for change in data analysis, the first model which should be fitted is the unconditional means model. This model provides a valuable baseline to be compared with the forthcoming models and partitions the total variation in the outcome meaningfully. The results of the unconditional means model help to establish whether there is systematic variation in the outcome that is worth exploring (Singer & Willett, 2003). Model 1 (Table 2) presents the results of fitting the unconditional means model to the teaching effectiveness. The average teaching effectiveness score is 89.973. The variance components indicates statistically significant variance associated with teachers (= 9.772, SE 3.77, p < 0.001) and statistically significant residual variance (= 19.03, SE 0.88, p < 0.001). Based on the calculation of the two estimated variance components, approximately 33.9% of the total variation can be explained by the individual difference or from differences among teachers. It can be concluded that teachers' teaching effectiveness varies over time and that teachers differ from each other in terms of teaching effectiveness. Because each variance component is significantly different than 0, it is possible to link both within-person and between-person variation in teaching effectiveness to predictor variables.

Model 2: Unconditional Growth Model

Model 2 is the unconditional growth model which partitions and quantifies the outcome variation across people and time and the results provide information on where that variation resides – within or between people (Singer & Willett, 2003). In this model, the fixed and random effects for time are incorporated into the level-1 sub-model and include no other predictors. The fixed effects for the initial starting point (intercept) and the slope of the population average change trajectory are both significant. Comparing the level-1 residual variance in Model 2 to that of Model 1, we find a decline of 0.22 (19.03-14.86)/19.03, in other words, entering the linear time effect decreases the level-1 variance by 22%. Thus, it can be concluded that 22% of the within-person variation in teaching effectiveness is systematically associated with linear time. This suggests more explanatory factors should be added for further analysis. The level-2 variance components are associated with the individual teacher growth parameters. The variance components showed statistically significant variance associated with teachers (=16.01, SE 3.77, p < 0.01) which means that individuals have different initial status. Residual variance is also (=14.86, SE 0.58, p < 0.01) statistically significant. The linear fixed effect of time differed significantly and positively from zero (= 0.34, SE 0.033, p < 0.01). The time related variance component indicated statistically significant, systematic differences among linear trends across teachers (=0.0319, SE 0.013, p < 0.001) indicating that individual teachers have different rate of change with time. The

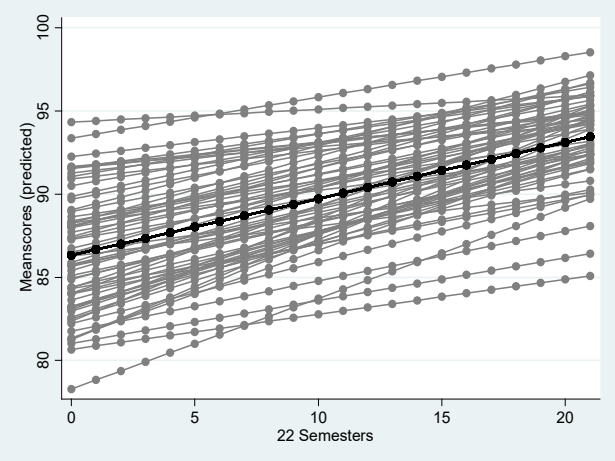
residual covariance component which assesses the relationship between the initial status and change of rate and it differed significantly from zero (= -0.508, SE 0.198, p < 0.01). Correlation coefficient of the relationship was estimated based on the variance components using the formula presented by Singer and Willett (2003):

$$\rho_{\pi 0 \pi 1} = \rho_{01} = \frac{\sigma_{01}}{\sqrt{\sigma_0^2 \sigma_1^2}} = \frac{-0.508}{\sqrt{(0.0319)(16.01)}} = -0.71$$

This shows a strong negative relationship between the initial status and the rates of change which means teachers who have lower starting points tend to have higher growth rate compared to teachers with higher starting points.

In Figure 1, the plots based on individual teachers from Model 2 demonstrate the amount of variation among teachers. Each of the 62 grey horizontal lines represents the linear effects of year for a different teacher. The bold line shows the average function across the 62 teachers. The graph shows that there is a substantial variation in the intercepts associated with each teacher which indicates the teachers have different initial status and that individual teachers have different rate of change with time.

Figure 1. Mean teacher ratings and ratings of 62 teacher



<u>Model 3</u>: Effect of Gender, Age, Academic Degree, Teaching Experience, Course Type, Number of Students, Semester

As suggested by the previous model, more explanatory factors are added in this model. In model 3, the predictor variables added to the model are gender, age, teacher's academic degree, years of teaching experience at the first measurement, course type, number of students and semester under the fixed effects part. As presented in Table 2, the three variables of course type, number of students, and semester have significant negative relationship with the student ratings of teaching effectiveness. It also shows that type of school has a small effect on the ratings which means that school which the teacher belongs contributes to the difference in the student ratings. The results indicate that the ratings for major course are higher than general course and ratings in fall semester are higher than in spring semester. As for the number of students, the negative relationship indicates that the more number of students that the teacher teaches, the lower teaching effectiveness score the teacher tends to receive.

Other predictors such as gender, age, teacher's academic degree and years of teaching experience all have no significant effects on teaching effectiveness. We also examined whether these variables have random effects but they have no significant effects. Adding these predictor variables slightly reduces level 1 within-person variance from 14.86 to 14.41. It also decreases the level-2 individual difference variance from 16.01 to 14.57. These variables explain 9% [(16.01-14.57)/16.01 of the between teacher variance.

<u>Model 4</u>: Effect of Course Type, Number of Students, Semester, School

Model 4 excludes all variables which have no significant relationship with teaching effectiveness and only includes statistically significant variables which are course type, number of students, semester and school which the teacher belongs. The model fit indices AIC (5666.952), BIC (5711.04) and deviance (5649.0) are lower compared among the other models. Thus, model 4 is selected as the final model and the detailed interpretation of the terms in the final model is provided below along with relevant figures for better illustration.

<u>Model 5 and 6</u>: Early Career Status

In Model 5 and Model 6 (Table 2), the early career variable as well as interaction between time and early career status were incorporated to examine whether the mean stability and the growth pattern for teachers early in their teaching career are different from teachers with more years of teaching experience. The literature on

teaching effectiveness indicates that early career teachers differ from their more experienced counterparts in terms of growth pattern as early career teachers follow a quadratic growth pattern (an initial increase followed by a subsequent decline) whereas more experienced and senior teachers show a gradual and linear decline. In this study, teachers with less than three years of experience were considered as early career teachers. The results in the following models show that the effect of early career status is statistically significant. Also, there is significant effect in terms of interaction between early career status and time which indicate that early career teachers showed more improvement compared to the more senior teachers. For one more year of teaching, the average rate of change in teaching effectiveness is 0.168 lower for non-early career teachers.

Model 7, 8, 9: Initial Status and Growth Rate

Model 7, 8 and 9 (Table 2) were constructed to examine the whether there is significant difference in the teachers' initial status in terms of their teaching effectiveness as judged by students as well as their changing rate across time. In the first stage, teachers were divided into three groups based on their mean score ratings in the first year of the study as average, below average and above average relating to their teaching effectiveness. Table 5 shows that there is significant difference in terms of teacher's initial status as well as their growth rate. Model 7 shows the initial status and the growth rate of below average teachers; Model 8 refers to average teachers and Model 9 refers to above average teachers respectively. In table 5, it can be seen that average teachers have the highest growth rate which is 0.397. Below average teachers improve 0.288 on the scale. Above average teachers improve 0.192 on the scale. It means that teachers with average ratings at the beginning of the study had improved most rapidly compared to the other teachers.

Interpretation of the Final Multilevel Model-Model 4

The results showed that certain predictor variables including course type (major and general), number of students who attended all the classes taught by each teacher in a given semester, and semester in which the course was provided had significant effects on the student ratings of teaching effectiveness. Other factors (gender, age, academic degree, years of teaching experience, school) are not included in the final model as they had no significant effects on the ratings.

Time (0.345): When all other variables are equal, the average teaching effectiveness scores increase by 0.345 (= 0.345, SE 0.033, $p < 0.01$) for every semester.

Course type (-1.468): The results indicate that course type has significant effect on student ratings of teaching effectiveness. Teachers who teach general courses are rated 1.468 score (= -1.468, SE 0.768, $p < 0.01$) lower than teachers who teach major courses. The following figure illustrates the difference in the ratings across 22 semesters by course type which the major courses are rated higher than general courses.

Number of students (-0.008): The results reveal the number of students enrolled in classes taught by the same teacher in one semester has significant impact on teaching effectiveness. The estimated teaching effectiveness score for teachers appears to drop 0.008 (= -0.008, SE 0.003, $p < 0.01$) on average when each additional student is added. The larger class size that the teacher teaches, the lower ratings score the teacher tends to receive.

Semester (-1.319): A significant difference in student ratings was found between two semesters in each academic year. The results show that teachers tend to receive ratings score 1.319 (= -1.319, SE 0.25, $p < 0.01$) lower in the second semester (spring semester) compared to the first semester (fall semester), when all other variables are equal. The following Figure 3 shows the difference of the teacher ratings between semesters across 22 semesters. Within each year there is a consistent difference between the teacher ratings in the first and the second semester in that teachers are rated higher in the first semester than in the second semester.

Check for model adequacy

Having fitted the model, we can predict best linear unbiased predictions of the teaching effectiveness. We examine quantile-quantile plot for the residuals along with the standardized residuals in the final model 4 to check whether the random effects are normally distributed. The quantile-quantile plot for the random effects shows that these effects are approximately normal. In addition, plots of standardized residuals show

that residuals are normally distributed. In order to save space, the figured are not provided here but we are ready to provide upon request.

Figure 2. Intercept Random Effects of 62 teachers

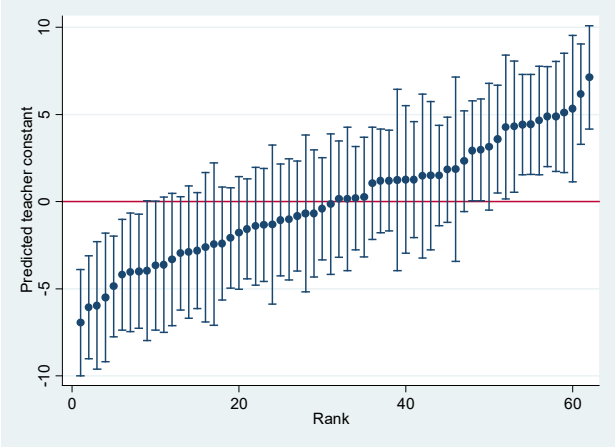


Figure 2 illustrates the amount of the intercept random effects of the final model with caterpillar plots. In the figure, all 62 teachers are ranked according to the teacher intercept from lowest to highest. Each vertical line on the graph represents the teaching effectiveness of individual teachers. The graph shows that there is substantial difference among teachers in terms of their teaching effectiveness.

Figure 3. Intercept Random Effects of 62 teachers

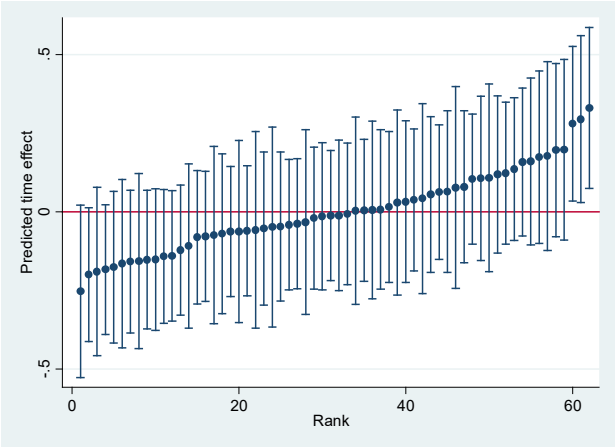


Figure 3 illustrates the amount of the time random effects of the final model with caterpillar plots. There are three teachers at the end of the graph for whom the 95% confidence interval does not include the mean growth across all teachers which these three teachers' intercept and growth rate were relatively higher than the other teachers

5. CONCLUSION

Multilevel longitudinal analyses using the dataset in this research show that the teaching effectiveness of a private university in Mongolia which was selected in this study improved significantly over an 11-year period. Teaching experience, age and academic degree which are three measures of seniority found to have no significant effect on the student ratings. This indicates that there are other reasons that have influenced the significant improvement in the teaching effectiveness. For example, changes in administration and administrative structure, accreditation process, financial incentives including reward for earning doctoral degree and salary increase for faculty members could have influenced the improved teaching effectiveness. Although the overall teaching effectiveness has improved, delivery skills of instructors in all disciplines has consistently received lower ratings which suggest that both the university and faculty need to take measures to improve this important teaching aspect[24].

The course type, number of students, and semester have significant negative relationship with the student ratings of teaching effectiveness. The results indicate that the ratings for major course are higher than general course and ratings in fall semester are higher than in spring semester which could have been influenced by ratings of freshman students as they constitute a large portion of all students participating in student ratings. For the number of students, the negative relationship indicates that the more number of students that the teacher teaches, the lower teaching effectiveness score the teacher tends to receive. Thus, the academic administrators need to take this matters into consideration when evaluating faculty from different disciplines and teaching different types of courses. The teaching load of instructors at this university needs to be assigned at appropriate level for all instructors in all disciplines as those who are assigned to teach multiple classes with large number of students are rated to be less efficient in their teaching.

The study findings also reveal that teaching effectiveness varies over time and that teachers

[24] In order to save space, this information is excluded from the article. It is available upon request.

differ from each other in terms of teaching effectiveness. There is a strong negative relationship between the initial status and the rates of change which means teachers who have lower starting points tend to have higher growth rate compared to teachers with higher starting points. Teachers have different initial status and that individual teachers have different rate of change with time. Teachers with average ratings at the beginning of the study improved most rapidly compared to those with high and low ratings. Early career teachers showed more improvement compared to more senior teachers.

Other predictors such as gender, age, teacher's academic degree, school type and years of teaching experience are all have no significant effects on teaching effectiveness.

Student ratings reflect students' satisfaction with instructor's teaching effectiveness which define teaching quality, consequently reputation of higher education institutions. They also provide valuable feedback for administrators and faculty members to assess their teaching performance and make an improvement in their teaching. With that in mind, the university and faculty members need to make an actual use of student ratings to examine which aspect of teaching needs improvement and take necessary actions that require dedication and effort from both university administration as well as faculty members.

Recommendations for Future Studies

More longitudinal studies need to be conducted to further explore about the student ratings in Mongolia as most previous studies in the field have been conducted in western countries especially in the US and other English-speaking countries. Considering the different social and cultural aspects, the results of those studies may not be generalized to countries such as Mongolia as some results of this study were not in line with previous studies conducted in western countries. More studies on student ratings need to be conducted in the context of developing countries.

This study examined the effects of variables related to teacher characteristics and course characteristics and further studies need to explore whether other factors such as salary, administrative and structural changes, and accreditation process could have influence on the teaching effectiveness. Moreover, further studies need to examine whether student related variables have effect on student ratings.

Due to the sensitivity of the data on student evaluation surveys, longitudinal data from multiple universities could not be accessible. Future studies are encouraged to conduct this type of study in different institutions of higher education with an extended sample size representing different disciplines.

In this study, mean rating scores of student ratings were used to determine whether certain instructor and course variables have impact on how students rate their teachers. Results may differ when rating scores of specific teaching dimensions are used to examine the effects of the variables. Therefore, future studies need to examine the impact of the variables on both the global ratings and ratings of specific aspects of teaching.

References:

Arreola, R. A. (1989). Defining and evaluating the elements of teaching. Proceedings of Academic Chairpersons: Evaluating Faculty, Students, and Programs (pp. 1-14). Manhattan: Kansas State University.

Arreola, R. A. (2000). Developing a comprehensive faculty evaluation system: A handbook for college faculty and administrators on designing and operating a comprehensive faculty evaluation system. Bolton, MA: Anker.

Baasandorj, Dolgorsuren. (2010). *Faculty development program needs at Mongolian state universities: Content and strategies.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 3413838)

Bat-Erdene, Mash-Ariun. (2006). *Faculty participation in decision-making and their job satisfaction in Mongolian public universities.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 3239789)

Bat-Erdene, R., Costa, V., & Yeager, J. (1996, March). *The impact on structural adjustment in the Ministry of Science and Education, Mongolia*. Paper presented at the Comparative and International Education

Society Conference, Williamsburg, Virginia.

Batrinchin, Pagma, Nerendoo Nergui, Damba Monkhor, Tsendjav Jargalmaa, Altangerel Hajidsuren, & John C. Weidman. (2002). *Mongolia Country Study*. RETA No. 5946-REG: Sub-Regional Cooperation in Managing Education Reforms. Manila, Philippines: ADB.

Batsukh, Tungalag. (2011). *Mongolian higher education reform during the transition to a democratic and market-based society 1990-2010.* Unpublished dissertation, University of Pittsburgh. Retrieved April 1, 2016 from http://d-scholarship.pitt.edu/10852/1/ETD_Batsukh_8_22_11.pdf

Bayartsetseg, Batjav. (2003). Багшийн сургалтын ажлыг социологийн аргаар үнэлэх аргазүй. Unpublished dissertation, Mongolian State University of Education, Ulaanbaatar, Mongolia.

Benton, S. L. & Cashin, W. E. (2014). Student ratings of instruction in college and university courses. *Higher Education: Handbook of Theory and Research, 29*, 279-325.

Bianchini, S., Lissoni, F., & Pezzoni, M. (2012). Instructor characteristics and students' evaluation of teaching effectiveness: Evidence from an Italian engineering school. *European Journal of Engineering Education,* 1-20. doi:10.1080/03043797.2012.7428

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited. IDEA Paper No. 32.* Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Carle, A. C. (2009). Evaluating college students' evaluations of a professor's teaching effectiveness across time and instruction mode (online vs. face-to-face) using multilevel growth modeling approach. Computers & Education, 53, 429-435.

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003). *Applied Multiple Regression/ Correlation Analysis for the Behavioral Sciences*, 3rd ed. Mahwah, NJ: Erlbaum.

d'Apollonia, S. & Abrami, P.C. (1997). Navigating student ratings of instruction. *American Psychologist*, *52*(11), 1198-1208.

Feldman, K.A., 1993. College students' view of male and female college teachers: part II, evidence from students' evaluations of their classroom teachers. *Research in Higher Education, 34 (*2), 151–211.

Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry, & J. C. Smart (Eds.). *The Scholarship of Teaching and Learning in Higher Education: An Evidence Based Perspective.* Dordrecht, The Netherlands: Springer.

Hallinger, P. (2010). Using faculty evaluation to improve teaching quality: A longitudinal case study of higher education in Southeast Asia. *Educational Assessment, Evaluation and Accountability, 22,* 253-274.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *The Journal of Economic Literature, 24,* 1141-1177.

Hativa, N. (2014, a). *Student Ratings of Instruction: Recognizing Effective Teaching.* Oron Publications, Nira@me.com

Hoyt, D. P. & Pallett, W. H. (1999). *Appraising teaching effectiveness: Beyond student ratings. IDEA Paper No. 36.* Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Huang, F. L. & Moon T. R. (2009). Is experience the best teacher? A multilevel analysis of teacher characteristics and student achievement in low performing schools. *Educational Assessment, Evaluation and Accountability*. 21 (3), 209-234. DOI: 10.1007/s11092-009-9074-2

Lang, J. W. B. & Kersting, M. (2007). Regular feedback from student ratings of instruction: Do college teachers improve their ratings in the long run? *Instruction Science, 35.* 187-205.

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. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, *76*, 707-754.

Marsh, H.W. (1987). Students' evaluation of university teaching: research findings, methodological issues, and directions for future /research. *International Journal of Educational Research, 11*(3), 263–388.

Marsh, H. W. (2007). 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., Overall J. U. & Kesler, S. P. (1979). Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. *Journal of Educational Psychology, 71*(2), 149-160.

Marsh, H. W., & Dunkin, M. (1992). Students' evaluations of university teaching: A multidimensional perspective. In G. A. Marcoulides, & R. E. Schumacker (Eds.), *Higher Education Handbook on Theory and Research* (8th ed., pp. 143-224). New York: Agathon.

Marsh, H. W., & Dunkin, M. J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 241–320). New York: Agathon.

Marsh, H. W. & Hattie, J. (2002). The relationship between research productivity and teaching effectiveness: Complementary, antagonistic, or independent constructs? *Journal of Higher Education, 64*(5), 1-17.

Marsh, H. & Roche, L.A. (1997). Making students' evaluation of teaching effectiveness effective: The critical issues of validity, bias and utility. *American Psychologist*, *52*(11), 1187-1197.

Ministry of Education, Culture and Science (2016). *Tertiary Education Statistics.* Retrieved April 1, 2016, from http://www.meds.gov.mn/HigherSta1516

Muijs, D. (2011). *Doing quantitative research in education with SPSS,* 2nd edition. Sage Publications.

Murray, H. G., Jelley, R. B., & Renaud, R. D. (1996). Longitudinal trends in student instructional ratings. Paper presented at annual meeting of the American Educational Research Association, New York.

Murray, H. G. (1997). Does evaluation of teaching lead to improvement of teaching? *International Journal for Academic Development*, *2*(1), 8-23. doi: 10.1080/1360144970020102

National University of Mongolia (2012). Faculty Evaluations Regulations. Retrieved April 1, 2016 from https://sisi.num.edu.mn/files/ZHSHUA/Juram/MUIS-iin%20bagshiin%20ajlyn%20negdsen%20unelgeenii%20(attestatchilalyn)%20juram%20.pdf

Seldin, P. (1984). *Changing practices in faculty evaluation: A critical assessment and recommendations for improvement*. San Francisco, CA: Jossey-Bass.

Singer, J. D. & Willett, J. B. (2003). Applied longitudinal data analysis: Modeling change and event occurrence. Oxford University Press.

Toch, T. & Rothman, R. (2008). *Education sector reports: Rush to judgement.* Retrieved April 1, 2016 from http://educationpolicy.air.org/sites/default/files/publications/RushToJudgment_ES_Jan08.pdf

APPENDIX
Table 2. Estimation results. Dependent variable: Overall teacher rating

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Time (Rate of change,)		0.340***	0.344***	0.345***	0.309***	0.371***	0.314***	0.345***	0.352***
		(0.033)	(0.040)	(0.033)	(0.039)	(0.048)	(0.040)	(0.033)	(0.034)
Semester			-1.320***	-1.319***	-2.110***	-1.322***	-1.313***	-1.319***	-1.324***
			(0.251)	(0.250)	(0.508)	(0.250)	(0.250)	(0.250)	(0.250)
Number of student			-0.008***	-0.008***	-0.008***	-0.005	-0.007***	-0.008***	-0.008***
			(0.003)	(0.003)	(0.003)	(0.005)	(0.003)	(0.003)	(0.003)
Course type			-1.628*	-1.468*	-1.476*	-1.456*	-2.676**	-1.504*	-1.501*
			(0.890)	(0.768)	(0.769)	(0.769)	(1.139)	(0.776)	(0.776)
Academic degree			0.141						
			(1.143)						
Teaching experience			-0.021						
			(0.100)						
Age			0.022						
			(0.087)						
Gender			-0.931						
			(0.847)						
school1			1.774*						
			(1.002)						
school2			1.203						
			(1.196)						
time*semester					0.074*				
					(0.041)				
time*studentnumber						-0.000			
						(0.000)			
time*coursetype							0.099		
							(0.071)		
Early								0.680	1.829
								(1.538)	(2.231)
time*early									-0.091
									(0.128)
Intercept (Initial status,)	89.973***	86.328***	87.008***	88.071***	88.444***	87.811***	88.417***	88.042***	87.983***
	(0.422)	(0.583)	(2.322)	(0.677)	(0.708)	(0.764)	(0.715)	(0.681)	(0.687)
Variance component									
Within-person	19.03***	14.86***	14.41***	14.41***	14.37***	14.42***	14.39***	14.41***	14.41***
	(0.88)	(0.58)	(0.699)	(0.699)	(0.697)	(0.7)	(0.697)	(0.698)	(0.699)
In rate of change		0.0319***	0.0331***	0.0329***	0.0332***	0.0323***	0.0338***	0.0330***	0.0333***
		(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.13)
In initial status	9.772***	16.01***	14.57***	14.53***	14.51***	14.53***	14.18***	14.52***	14.59***
	(2.01)	(3.77)	(3.76)	(3.54)	(3.54)	(3.55)	(3.45)	(3.56)	(3.60)
Covariance		-0.508***	-0.479***	-0.484***	-0.484***	-0.48***	-0.477***	-0.479***	-0.484***
		(0.198)	(0.194)	(0.189)	(0.189)	(0.189)	(0.186)	(0.189)	(0.193)
N	991	991	991	991	991	991	991	991	991
AIC	5872.962	5685.554	5674.994	5666.952	5670.292	5682.147	5670.486	5666.065	5669.834
BIC	5887.658	5714.946	5748.475	5711.040	5719.279	5731.134	5719.473	5715.052	5723.720
Deviance	5867.0	5673.6	5645.0	5649.0	5650.3	5662.1	5650.5	5646.1	1000.1

Standard errors in parentheses; * p<.1 ** p<.05 *** p<.01