Basic Statistical Analysis and Modelling of Evaluation Data for Teaching

A Master Thesis Presented

by

Yilan Zhou
(176397)

Tester: Prof.Dr.B.Rönz
Director: Dr.S.Klinke

CASE-Center for Applied Statistics and Economics

Institute for Statistics and Econometrics

in partial fulfillment of the requirements

for the degree of

Master of Arts

Humboldt-Universität zu Berlin

School of Business and Economics

Spandauer Str. 1

D-10178 Berlin

Berlin, July 6, 2004

Declaration of Authorship

I hereby confirm that I have authored this master thesis independently and without use of others than the indicated resources. All passages, which are literally or in general matter taken out of publications or other resources, are marked as such.

Yilan Zhou

Berlin, 6th July 2004

Abstract

This thesis proposes a novel numerical scoring system, which efficiently evaluates the teaching effectiveness of the lecturers. Based upon the scores given in the student evaluation of teaching (SET), this numerical scoring system employes the *factor score of one-factor model of data* and yields the instructor rankings result as output.

The other purpose of this paper is to discover determinants of SET scores, especially to examine whether factors which are normatively irrelevant to teaching quality matter or not. Results indicate that communication skill of lecturer & students' reaction, course attributes and quality of lecture notes are three most significant factors which determines the student response to "general overall ratings" of the course. The study suggests that class size and class meeting time also have some influence on that.

Keywords: Chi-square statistics, Corrected Contingency Coefficient, Normalized Uncertainty Coefficient, Underlying Variable Approach, Multinomial Logit Model

Acknowledgments

First of all, I would like to appreciate my supervisor, Dr. Sigbert Klinke, for his highly responsible supervision, his unfailing encouragement, and his invaluable guidance undertaking this study.

My biggest and deepest gratitude goes to Prof.Dr.Bernd Rönz and Prof.Dr.Wolfgang Härdle for instructing me to the new area of statistics of social science and giving me inspiration and courage to explore it.

I am also grateful to my dear friends, Qingfan Jia, Jeffery Tao, and Ying Chen, and my family, my father, Xianning Zhou and my mother, Qingfen Li, for their warmest encouragement.

Contents

List of Tables

List of Figures

1 Introduction

Student evaluation of teaching (SET) is widely used in tertiary institutions to measure instructor performance and to further improve the course quality. The evaluation office in the School of Business and Economics, Humboldt Universität zu Berlin examines the effectiveness of teaching based upon the data compiled from the course evaluation forms that are distributed to students each semester. To retrieve the information accurately from and make correct interpretation of the data, advanced statistic analysis must be carefully selected and properly applied on the collected data.

Currently, the mean scores of all the numeric items in the evaluation form (e.g., the global overall ratings) are calculated and then used as a major indication of the effectiveness of the teaching in the report prepared by the evaluation office, refer to Evaluation (2002) and Evaluation (2003). Nevertheless, there is no de-facto standard approach defined so far to measure the "teaching ability" in the general sense, which unfortunately incurs unnecessary ambiguity and significant inconsistency, when the effective/ineffective instructors are to be identified. Given the tremendous emphasis that the university places on teaching excellence in its annual merit review and in its promotion policy and tenure selection process, a quantified standard criterion becomes indispensable. In this thesis, a practical method is proposed to find the course that has the best teaching quality. Specifically, using factor analysis on the students' ratings, a single indicator, which closedly reflects instructors' general teaching ability, can be identified. The ranking of teaching effectiveness of each instructor then can be determined, based upon the mean scores of the single indicator.

Our finding also sheds light on possible sources of student evaluations. Still by factor analysis on SET, we recognize five separate dimensions of instructional effectiveness, namely lecturer's communication skill, quality of lecture notes, course attributes, students reactions and question answering. But as a complex multidimensional activity, teaching also comprises of a number of a separable variables such as teacher's characteristics (e.g., gender, reputation), course characteristics (e.g., meeting time, class size) and students' characteristic (e.g., gender, major). SET instruments should also reflect this multidimensionality. By employing multinomial logic regression techniques, we find that it measures not only aspects of instructional effectiveness, but also captures some factors that are normatively considered irrelevant to teaching quality.

During the analysis on the data obtained from the evaluation form, we have also indentified some problems existing in the structure of form and provided some suggestions about the improvement.

The data set is first overviewed in next section. The main statistical methods used in study are described in section 3. Results of factor analysis will be interpreted in section 4. The numerical scoring system will be introduced in section 5 and outcomes of multinomial logit model of data will be presented in section 6. Conclusions will be drawn and potential development will be discussed in the final section. The main softwares used in this work are XploRe, M-plus, SPSS 11.0.

2 General Overview of Data

2.1 Data Overview

The data used in the study is extracted from the questionnaires in the evaluation form, which is distributed to students each semester by the evaluation office in the School of Business and Economics, Humboldt Universität zu Berlin. Three types of forms have been designed, each of which specifically targets the lecture course, exercise course and seminar course, respectively. Since the content of questionnaires and structure of the form used in the seminar course is totally different from those of the other two types of courses, we choose to focus on the evaluation data for lecture course and exercise course and analyze them separately.

The questionnaire contains six sections, see in Figure A.1: The first section collects students information such as gender, major, course miss times, reasons why students miss, and global overall ratings of the course. The other five sections include thirty three general response items, which concentrate on specific aspects of teaching, e.g., lecturer, lecture concept, course attributes, self assessment of students, and course atmosphere. Each item uses a five point scale, ranging from 1(very good or too high) to 5(very bad or too low). The reverse side of the form includes 4 item, which asks for verbal comments on the strongest points of the course, the weakest points of the course, the suggestions on future improvement and other constructive comments on the course such as the room size. This paper only deals with the numeric items.

For sake of completeness, refer to information materials of the courses, e.g., *Studienordnung für den Diplomstudiengang Betriebswirtschaftslehre 2000, Humboldt Universität zu Berlin* (2000) and *Studienordnung für den Diplomstudiengang Volkswirtschaftslehre 2000, Humboldt Universität zu Berlin* (2000), following pieces of information for each course have been introduced as variables in the quantitative analysis hereafter:

- Class size: number of the students in the class.
- Class time: the time of day the course meet.(before 2:00 pm is morning class; after 2:00 pm is afternoon class)
- Day of class: the day the course meets.(on the border of week: Monday or Friday; in the middle of week: Tuesday till Thursday.)
- Class level: undergraduate class or graduate class.
- Class compulsory: compulsory for student or not.
- Instructor's gender: male or female.
- Instructor's rank: professor or assistant.

The data set used in this study covers two summer academic semester of 2002 and 2003 and consists of one hundred and sixty four individual undergraduate and graduate courses taught by more than thirty five instructors. For illustration purpose, hereby, the data sample for lecture course 2003 will be discussed in more detail.

There are over 10500 response observations in the whole four datas, which comprises students mainly majoring in economics (VWL) and

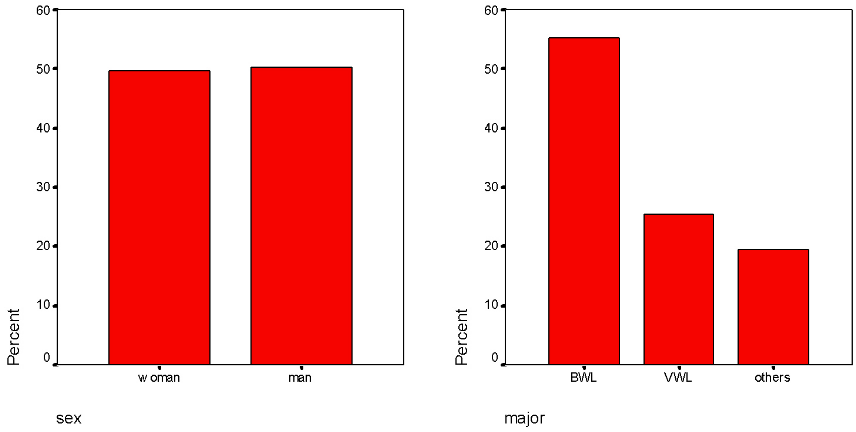


Figure 2.1: Distribution of student, left: gender, right: major.

management (BWL). It is noted from Figure 2.1, which plots the distribution of students major and gender, that more management students are integrated in the data, and male students and female students each occupy about 50 percent of data.

The data set includes both the courses taught by professors and the ones whose instructors are assistants. Figure 2.2 reveals that among the instructors who teach lecture course, there are over 80 percent of males with the rank professor. In the meantime, almost all of the exercise courses are delivered by assistant teachers, which is confirmed by the statistics listed in Table A.2. It is also worthwhile to note that the number of female instructors increase in exercise class.

In German education system, all courses designed for bachelor students are mandatory. Only master students have chances to choose courses which are optional. As portrayed in Figure 2.3 and Figure 2.4, the data to be analyzed contains more mandatory courses in undergraduate level than optional ones.

The courses usually meet from 8*a.m.* till 8*p.m.*, Monday through Fri-

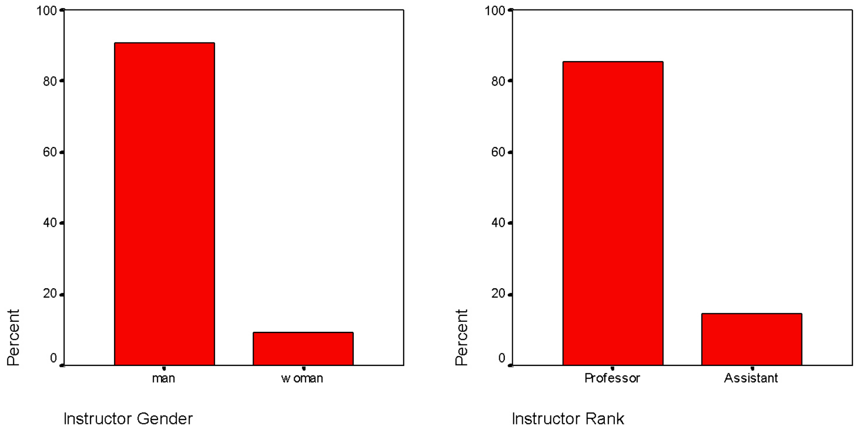


Figure 2.2: Distribution of teacher, left: gender, right: rank.

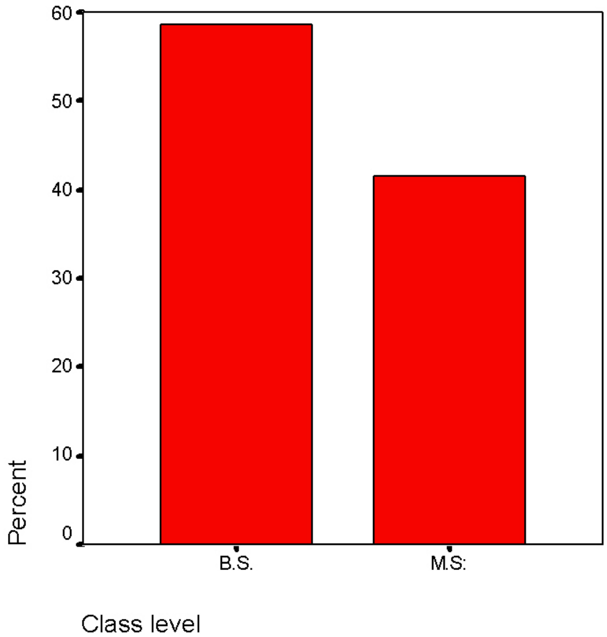


Figure 2.3: Class level, left: undergraduate, right: graduate.

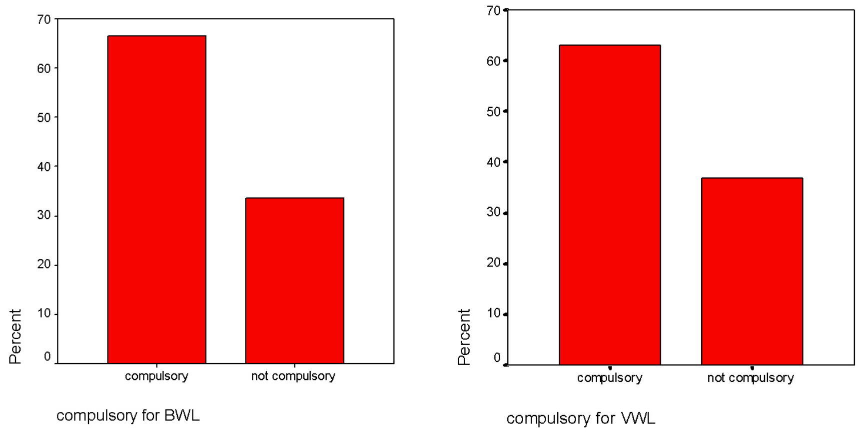


Figure 2.4: Class compulsory, left: BWL, right: VWL.

day. Figure 2.5 depicts the distribution of class meeting time. It can be observed that more classes are scheduled in the morning before $2p.m.$.

The size of each each individual class ranges from less than 10 to over 300, which is shown in Table A.1. Normally complusory course in undergraduate level has a big class size, over hundreds of students, see in Figure 5.2.

Table A.3 - Table A.5 have shown the detailed distributions of all 4 datasets.

2.2 Missing Value Imputation

No perfect data exists in the real world. Missing values in the data set always present significant problems in statistical analysis. The Table A.6 illustrates the percentages of missing data in each item variable. It is obvious that the missing data must be properly handled before any serious statistic analysis.

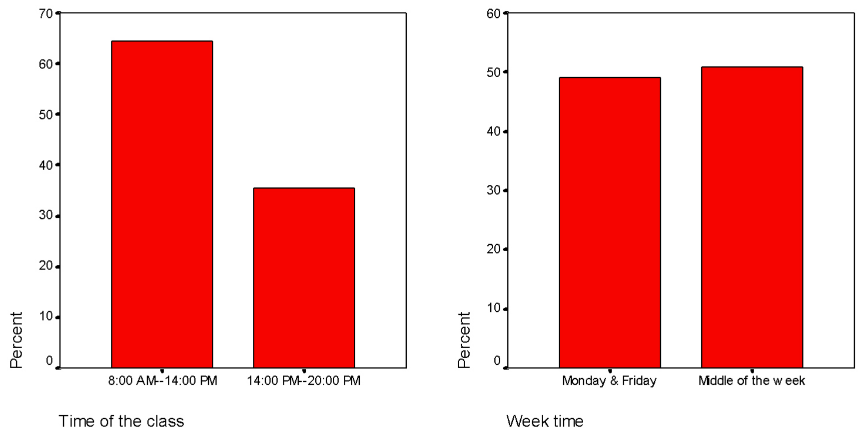


Figure 2.5: Class meeting time, left: day time, right: week time.

One simple approach to dealing with the incomplete data is to drop the corresponding observations. Easy as it is, this method however is the major culprit of potential inacurracy, especially when the sample population is small. Specifically, the analytical sample size will be reduced and precision of the evaluation be degraded, if the missing data discarded is correlated with the quantities of interest.

Therefore, we propose a systematic approach to fill in the missing data, which is described as follows:

1. Add one new category for missing values to item variables with high percentage missing values. From Table A.6,we can find high percentage missing values in some items such as "global overall ratings of the course"(over 10 percent); "course missing reasons"(over 40 percent); "time allowed after class"(over 25 percent); "relevance between lecture and exercise"(over 30 percent); "challenging feeling"(over 10 percent). Dealing with these items we recode missing value of these items into 0 with the assumption that students not willing answer the question. This way of

imputation has its advantage, not losing too much information. On the other hand, it makes the value of some item variables not ordered any more, leading bias to the data set.

2. Impute the data of other variables with small percentage missing values. There are a lot of popular imputation methods for categorical data, such as multiple imputation (MI), refer to Schafer & Olsen (1997). For the reason that it is too complex to programme in XploRe, Dr. S. Klinke suggests to use two imputation methods here:

- Mode substitution.
- Conditional mode substitution.

The idea of mode substitution is to replace every missing data point with mode of valid data for the variable. It sounds like a reasonable method, but as the same value is being substituted to each missing case, this method artificially has reduced the variance of the data and seriously dampened the relationship among variables.

Conditional mode substitution is treating the missing value as the dependent variable to be estimated using the data that exist. Suppose there are p variables in the dataset, and we want to impute the missing values variable k in the jth observation. First from the comparison of the corrected contingency coefficient (CC) between the vaiable k and other variables i, $i = 1, ..., p$, we pick out the variable m which has the highest CC with k,

$$m = \{l|C_{lk} = max_{i=1,...,p,i\neq k}C_{ik}\} \tag{2.1}$$

Where C_{ik} is the corrected contingency coefficient between i and k. Second, suppose, corresponding to missing value of variable k in the

	Number of Missings	Number of differences
Lecture2002	1858	762
Exercise2002	954	324
Lecture2003	1319	534
Exercise2003	752	272

Table 2.1: Comparison of Imputation methods.

matrixcomp.xpl

jth observation, the value of variable m of jth observation is v_{jm}, find the conditional mode of k, v_{jk}, the value of variable k which occurs most often when $V_m = v_{jm}$ to fill in the missing value of k, presented as following equation:

$$v_{jk} = mode(V_j | V_m = v_{jm}) \quad (2.2)$$

But if the value of v_{jm} is also missing, we will choose to impute the missing values in variable m first and then variable k. Iterations process is used in this method.

After imputation, we can see the difference between two imputation methods, shown in Table 2.1. We have chosen the second method to utilize the information that other variables could lend.

The XploRe prgramms of imputations are listed in attached CD (directory:appendix/xplore).

2.3 Descriptive Analysis of Response Data

Before we make any advanced statistical analysis on the response data, it is necessary to explore the response patterns of students. For data's description, we have calculated the frequency of selection, mode, and normalized entropy of response data to each item. The result of all four datasets are shown in Tables A.7 - A.10. From these tables, we can find some points very interesting about the data.

2.3.1 Skewness

It has been widely observed from the Tables A.7 - A.10 that the responses skew with most ratings at the positive side of the scale. Of note is that only about 15 percent of item are responded "bad" and "very bad" in each data set. There are at least two possible reasons for this event. One is that most instructional experiences may in fact be very good. Another is that students are always unwilling to give very bad ratings.

2.3.2 Entropy

Entropy coefficient shows the variability of the response. From Figure 2.6 and Figure 2.7, which plot the distribution of students response patterns to item variables with different value of entropy coefficient, we can find that the response of variables with small values of entropy are more concentrated around mode and those with big values are distributed more dispersely.

Comfirmed by the list of entropy coefficient list in Tables A.7 - A.10, students' responses to general features of the course, such as "Global overall ratings of the course", "mathematical level" and "difficulty level", do not differ much. But at the same time, to special characteristics with respect to the teaching quality, students' reactions are not alike.

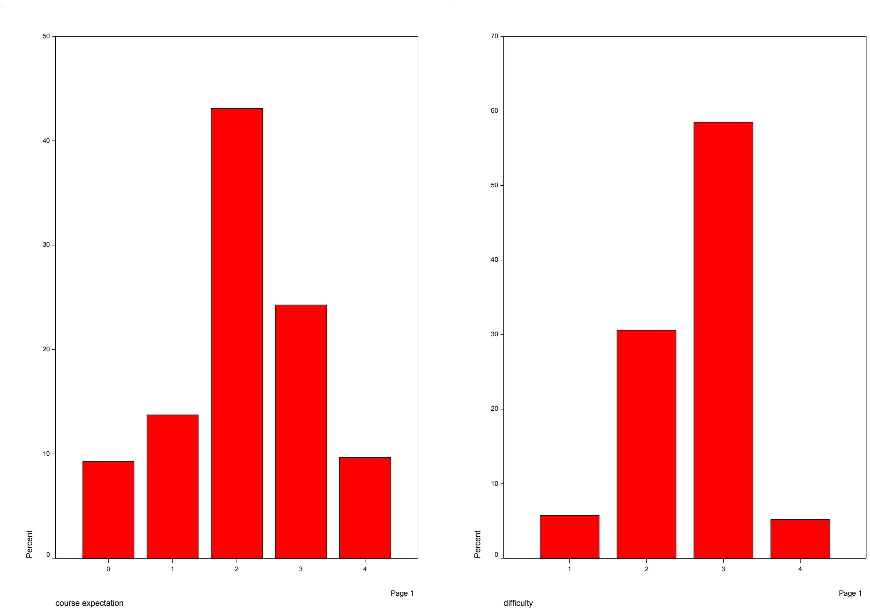


Figure 2.6: Entropy 0.18, left mode = 2, right mode = 3.

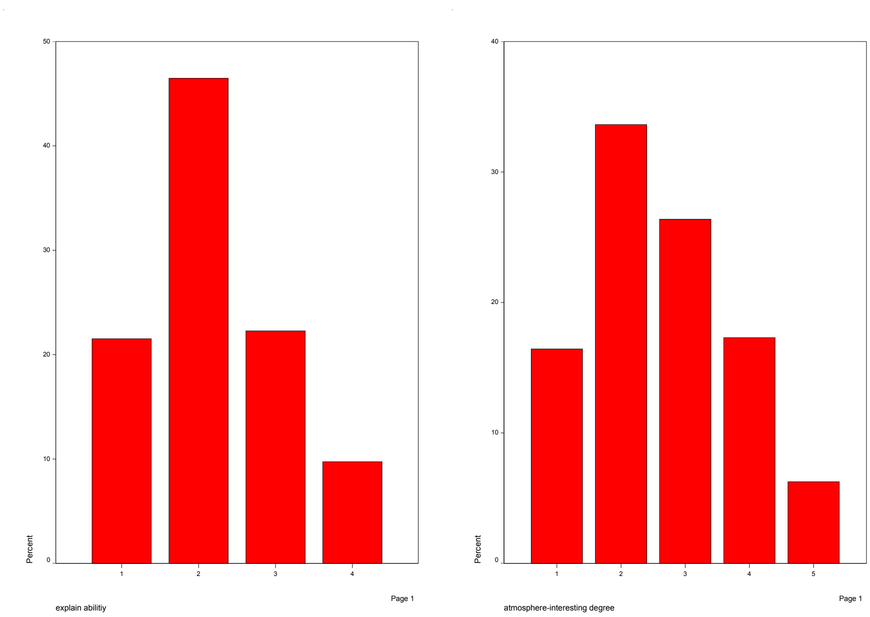


Figure 2.7: Entropy 0.26, left mode = 2, right mode = 3.

	0	1	2	3	4	5
Lecture course 2003	42%	5%	7%	5%	7%	34%
Exercise course 2003	54%	3%	5%	1%	6%	30%
Lecture course 2002	41%	6%	5%	6%	7%	34%
Exercise course 2002	55%	3%	5%	2%	5%	31%

Table 2.2: Frequency table for item "course missing reason".

2.3.3 Comparison of Students' Attitudes

Students' responses vary across their major and gender. Picking up one comparatively good course and one course with a relatively low teaching effectiveness from data sample, we have found some points worthwhile to note, although here general confirmation cannot be made just based upon only two course sample.

Figures 2.8 and 2.9 compare the students' expectation (Global overall ratings of the course) to the course according to their major and gender. When students met with a bad course, see Figure 2.8, economic students are not as critical as management students. They are not likely to give extreme bad ratings. On the other hand, when students met with a good course, see Figure 2.9, management students' ratings are highly concentrated in the good level and the ratings of economic students tend to be moderate. In some senses, we can say the management students are more sensitive to the quality of teaching. In the mean time, considering the gender of students, we have found that women are more willing to criticize than men.

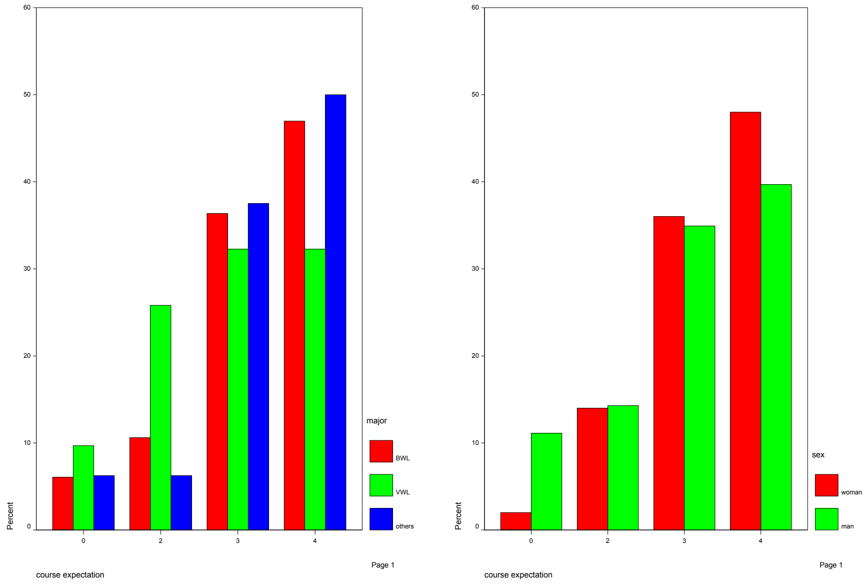


Figure 2.8: Bad course, left: major, right: sex.

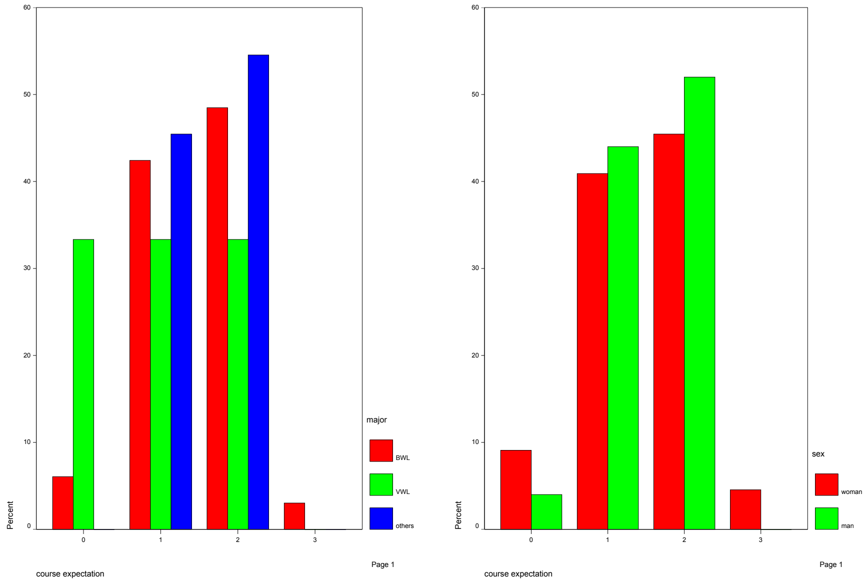


Figure 2.9: Good course, left: major, right: sex.

2.3.4 Discussion of Item "Course Missing Reason"

Another point here is, looking at the item "course missing reason", from the Table 2.2, we can see in both 2002 and 2003, wherever in the evaluation form for lecture course or exercise course, over 80 percent of item response were "0", which stands for missing data, or "5", which means "other reasons". That appears nearly 80 percent of information is not known, which means this item has no sense in the form. And we would suggest that this item can be omitted in the new form.

3 Statistic Methods

Before we make further advanced statistical analysis, we will first have a short look at concepts and ideas of the statistical method we carried out in this paper.

3.1 Univariate Analysis for Categorical Data

3.1.1 Mode

The mode x_{mod} of a set of numbers is the one that occurs most often. The formula is follows:

$$x_{mod} = \left\{ x_j | f_j = \max_{x_k} f_k \right\} \tag{3.1}$$

where f_k is the frequency which x_k appears. When more than one value occurs with the same greatest frequency, each value is a mode.

3.1.2 Entropy Statistics

Entropy is one measure to uncertainty of categorical data, similar to the variance, which measures the spread of random variables. The difference between these two concepts is that entropy applies to qualitative rather than quantitative values, and depends exclusively on the probabilities of possible events.

Let A_i stand for an event and $f(A_i)$ for the probability of event A_i to occur. Let there be N events $A_1, ..., A_N$ with probabilities $f(A_1), ..., f(A_N)$ adding up to 1. Entropy H can be computed by the following formula:

$$H = -\sum_{i=1}^{N} f(A_i) \ln f(A_i) \tag{3.2}$$

Normalized entropy are more often used because it ranges from 0 to 1.

$$H_0 = \frac{H}{H_{max}} = \frac{H}{\ln n} \tag{3.3}$$

The normalized entropy value is 0 corresponding to the case in which one event has unit probability. When all states are equally probable, it reaches to its maximum 1.

3.2 Bivariate Analysis for Categorical Data

3.2.1 Pearson Chi-squared Tests of Independence

Suppose two categorical response variables X and Y, X has J levels ($j = 1, ..., J$) and Y has K levels ($k = 1, ..., K$). The cells in contingency table represent the number h_{jk} of observations that $(X = j, Y = k)$. The hypotheses for the independence of them are:

H_0 : The X and Y are independent, i.e., $f(X = j, Y = k) = f(X = j) \cdot f(Y = k)$ for every pair (j, k).

H_1 : The X and Y are not independent, i.e., $f(X = j, Y = k) \neq f(X = j) \cdot f(Y = k)$ for every pair (j, k).

The Pearson chi-squared statistic for testing H_0 is

$$V = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(h_{jk} - \hat{e}_{jk})^2}{\hat{e}_{jk}} \tag{3.4}$$

where

- h_{jk} is the observed absolute frequency.
- $\hat{e}_{jk}$ is the expected absolute frequency calculated under the assumption that the two variables are independent when sample size is n.

$$\hat{e}_{jk} = \frac{h_{j+}h_{+k}}{n} \tag{3.5}$$

i.e., h_{j+} is the number of oberservations in the condition that $(X = j)$ and h_{+k} is the number of oberservations when $(Y = k)$.

Under H_0 the V statistic has approximately a χ^2 distribution for large sample sizes, with the degree of freedom $DF = (J - 1)(K - 1)$. In order to make test statistics follow the χ^2 distribution, the following two conditions should be satisfied when we are doing the test:

1. The estimated expected frequency $\hat{e}_{jk}$ of every cell should be larger than 1.
2. At most 20% of estimated expected frequency $\hat{e}_{jk}$ is smaller than 5.

The critical value $c = \chi^2_{1-\alpha;DF}$ for $f(V \leq c) = 1 - \alpha$, where α is the significance level. The null hypothese will be rejected when $v > \chi^2_{1-\alpha;DF}$.

It should be noted that the chi-squared test is quite sensitive to the sample size. The chi-squared value is overestimated if the sample size is too small and underestimated vice versa. To overcome this problem, contingency coefficient is one of the measures of association are often used.

To see the details, please refer to Rönz (1997).

3.2.2 Contingency Coefficient

The coefficient of contingency is a Chi-square-based measure of the relation between two categorical variables. It is computed by the following formula:

$$K = \sqrt{\frac{\chi^2}{\chi^2 + n}} \tag{3.6}$$

Where χ^2 is calculated by formula 3.4. Its value is between 0 and K_{max} where

$$K_{max} = \sqrt{\frac{M-1}{M}}, M = min\{J, K\} \tag{3.7}$$

The corrected contingency coefficient is:

$$K^* = \frac{K}{K_{max}} \tag{3.8}$$

Becuse the range of corrected contingency coefficient is always limited from 0 to 1 , where 0 means complete independence, it has advantage over the ordinary Chi-square is that it is more easily interpreted.

3.2.3 Uncertainty Coefficient

Uncertainty Coefficient(UC), which is also called entropy coefficient, varies from 0 to 1. From contingency table, the entropy of variable

X, can be computed by

$$U_X = -\sum_{j=1}^{J} f_{j+} ln f_{j+} \tag{3.9}$$

$$\tag{3.10}$$

for variable Y

$$U_Y = -\sum_{k=1}^{K} f_{+k} ln f_{+k} \tag{3.11}$$

for both variable X and Y

$$U_{XY} = -\sum_{j=1}^{J} \sum_{k=1}^{K} f_{jk} ln f_{jk} \tag{3.12}$$

The formula for $U_{(X|Y)}$, which is the uncertainty coefficient for predicting the row variable on the basis of the column variable, is given below as

$$U_{(X|Y)} = \frac{U_X + U_Y + U_{XY}}{U_Y} \tag{3.13}$$

Symmetrically, $U_{(Y|X)}$, which is the uncertainty coefficient for predicting the column variable on the basis of the row variable, is

$$U_{(Y|X)} = \frac{U_X + U_Y + U_{XY}}{U_X} \tag{3.14}$$

And symmetric uncertainty coefficient is

$$U = 2\left(\frac{U_X + U_Y - U_{XY}}{U_X + U_Y}\right) \tag{3.15}$$

The uncertainty coefficient is the percent reduction in uncertainty in predicting the dependent variable based on knowing the independent variable. When UC is 0, the independent variable is of no help in predicting the dependent variable. More detailed discussion about UC, see Rönz (1997).

This is to be contrasted with Pearsons correlation coefficient r_{xy}, which measures only linear correlation between two variables, i.e.,

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2 \sum\limits_{i=1}^{n}(y_i - \overline{y})^2}}$$

When the correlation is squared (r_{xy}^2), we get a measure of how much of the variability in one variable can be "explained by" variation in the other.

3.3 Kernel Density Estimation

The purpose of density estimation is to approximate the probability density function f of a random variable X. Assume there are n independent observations $x_1, ..., x_n$ from the random variable X. The kernel density estimator $\hat{f}_h(x)$ for the estimation of the density value $f(x)$ at point x is defined as

$$\hat{f}_h(x) = \frac{1}{nh}\sum\nolimits_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \tag{3.16}$$

where K is kernel function and h denoting the bandwidth.

For computation, the kernel function K must be evaluated to $O(h \cdot n^2)$ times, and the computation time will be increased if the sample size n is large. In practice, for graphing the density estimate, it is not necessary to calculate the $\hat{f}_h(x)$ for all observations $x_1, ..., x_n$. The estimate can be computed for example on an equidistant grid $v_1, ..., v_m$:

$$v_k = x_{min} + \frac{k}{m}(x_{max} - x_{min}), k = 1, ..., m << n \tag{3.17}$$

The evaluation of density requires then only $O(h{\cdot}n{\cdot}m)$ steps. This paper approximate the kernel density estimate by the WARPing method, refer to Härdle, Klinke & Müller (2001).

3.4 Exploratory Factor Analysis for Ordered Categorical Variables

3.4.1 Standard Factor Analysis

Factor analysis is a model-based technique to express the regression relationship between manifest variables $x_1, x_2, ..., x_p$ and latent variables $y_1, y_2, ..., y_q$. It aims to identify a set of latent variables $y_1, y_2, ..., y_q$, fewer in number than the observed variables($q < p$), that represent essentially the same information. When observed variables are metrical, the general linear factor model takes the form:

$$x_i = \alpha_{i0} + \alpha_{i1}y_1 + \alpha_{i2}y_2 + ... + \alpha_{iq}y_q + e_i \, (i = 1, ..., p) \qquad (3.18)$$

where $y_1, y_2, ..., y_q$ are common factors, e_i are residuals, and $\alpha_{i1}, ..., \alpha_{iq}$ are called loadings. Assumptions of the model are:

1. $y_1, y_2, ..., y_q$ are uncorrelated, and each has mean of zero and variance of one.
2. $e_1, e_2, ..., e_p$ are uncorrelated to each other, and each has mean of zero and variance. $Var(e_i) = \sigma_i^2.(i = 1, ..., p)$.
3. the ys are uncorrelated with the es.

The maximum likelihood method and the principal component method are used to estimate the standard factor model, which are discussed by Härdle & Simar (2003).

Factor scores, the estimated values of the factors, are also useful in the interpretations. The regression method to estimate is the simplest technique to implement, the details of this method introduced in Härdle & Simar (2003).

3.4.2 Exploratory Factor Analysis for Ordered Categorical Variables

When observed variables $x_1, x_2, ..., x_p$ are categorical, our object instead is to specify the probability of each reponse pattern as a functions of latent variables $y_1, y_2, ..., y_q$, takes the form

$$P(x_1 = a_1, x_2 = a_2, ..., x_p = a_p | y_1, y_2, ..., y_q) = f(y_1, y_2, ..., y_q) \quad (3.19)$$

where $a_1, ..., a_p$ represent the different response categories of $x_1, ..., x_p$, respectively, $f(y_1, y_2, ..., y_q)$ is a kind of function of latent variables $y_1, y_2, ..., y_q$.

Two approaches are often used in factor analysis for ordered categorical data: The underlying variable approach(UV) and item response function approach(IRF). For the reason that the former althogirithm is used in M-plus, the software which we used in our analysis, here we first give detailed description to the underlying varialbe approach and then compared it with IRF approach.

The Underlying Variable Approach

The underlying variable approach (UV) is similiar in spirit to factor analysis.In UV approach, We suppose each categorical variable x_i is generated by an underlying unobserved continuous variable x_i^* which is normally distributed with mean μ_i and variance σ_i^2.

The connection between x_i and x_i^* is that: for variable x_i with m_i categories, there are $m_i - 1$ threshold parameters: $\tau_{i(1)}$,$\tau_{i(2)}$,...,$\tau_{i(m_i-1)}$, then

$$x_i = s \Leftrightarrow \tau_{i(s-1)} < x_i^* < \tau_{i(s)}, (s = 1, 2, ..., m_i)$$

where

$$\tau_{i(0)} = -\infty, \tau_{i(1)} < \tau_{i(2)} < ... < \tau_{i(m_i-1)}, \tau_{i(m)} = +\infty$$

The model takes the form:

$$x_i^* = \alpha_{i1}^* y_1 + \alpha_{i2}^* y_2 + ... + \alpha_{iq}^* y_q + e_i \tag{3.20}$$

under assumptions:

- The latent variables y_i are independent and normally distributed with mean 0 and variance 1.
- The residuals are independent and normally distributed with mean 0 and variance σ_i^2.
- Univariate and bivariate normality of the underlying variables x_i^*.

By estimating the correlations between the underlying variables, x_i^*, which is also called the polychoric correlations, we have carried out a standard factor analysis.

In order to fit the model, three different sets of parameters: the thresholds, the polychoric correlations, and factor loadings of equation 3.20 are to be estimated. M-plus use three-step procedures, see in Muthén (1998):

- Thresholds are estimated from the univariate margins of the observed variables.

- Polychoric correlations are estimated from the bivariate margins of the observed variables for given thresholds.
- The factor model is estimated from the polychoric correlations by weighted least squares using a weight matrix.

IRF Approaches and its Relationship with the UV Approach

IRF approach specifies the conditional distribution of response pattern as a function of the latent variables. Let us suppose that there are m_i category for response variable i labelled $(1, ..., m_i)$; $\pi_{i(s)}(y)$ be the probability that, given y, a response falls in category s for variable i. Taking into account the ordinality property of the items we model the cumulative response probabilities,

$$\gamma_{i(s)}(y) = P(x_i \leq s) = \pi_{i(1)}(y) + \pi_{i(2)}(y) + ... + \pi_{i(s)}(y) \qquad (3.21)$$

and

$$1 - \gamma_{i(s)}(y) = P(x_i\ s) = \pi_{i(s+1)}(y) + \pi_{i(s+2)}(y) + ... + \pi_{i(m_i)}(y) (3.22)$$

where x_i stands for the category into which the ith variables falls.

The response category probabilities are denoted by

$$\pi_{i(s)}(y) = \gamma_{i(s)}(y) - \gamma_{i(s-1)}(y) \qquad (3.23)$$

The model used is the proportional odds model

$$log\left[\frac{\gamma_{i(s)}(y)}{1 - \gamma_{i(s)}(y)}\right] = \alpha_{is} + \sum_{j=1}^{q} \alpha_{ij} y_j; \qquad (3.24)$$

where $(s = 1, ..., m_i; i = 1, ..., p).'$

The assumptions:

- The latent variables are independent and normally distributed with mean zero and variance one.
- The responses to the ordinal items are conditional independent on the latent variables.

The above two methods of factor analysis of categorical data, UV and IRF, are discussed in detain in chapter 7 and 8, J.Bartholomew, Steele, Moustaki & I.Galbraith (2002).

Though the UV and the IRF models look quite different in model fitting procedure and assumption, but the equivalence has been noticed between 2 methods by Bartholomew and Knott(1999) and described in J.Bartholomew et al. (2002). The equivalence in the general case is showing the following relationships between the parameters of the two models:

$$\alpha_{ij} = \frac{\alpha^*_{ij}}{\sigma_i} \tag{3.25}$$

$$\alpha_{i(s)} = \frac{\tau_{i(s)}}{\sigma_i} \tag{3.26}$$

where $\tau_{i(s)}$ is the thresholds, α^*_{ij} is the factor loading of the jth latent variable and σ^2_i is the variance of the error term in the linear factor model for ith ordinal variable. For the factor analysis model of equation 3.20, the correlation between a underlying variable x^*_i and a latent variable y_i is

$$Corr(x^*_i, y_i) = \frac{\alpha^*_{ij}}{\sqrt{\sum_{j=1}^{q} \alpha^{*2}_{ij} + \sigma^2_i}} \tag{3.27}$$

Replace 3.25 into 3.27, the same correlation in terms of the IRF parameter α_{ij} will be got. And standardized value of α_{ij}, $st\alpha_{ij}$ takes the form:

$$st\alpha_{ij} = \frac{\alpha_{ij}}{\sqrt{\sum_{j=1}^{q} \alpha_{ij}^{*2} + 1}} \tag{3.28}$$

Although IRF is preferred because it makes use of the full distribution over all the other patterns, for the reason that these two methods get the same result, it will not affect the result of our analysis that we choose to use UV.

3.5 Multinomial Logistic Models

Multinomial Logistic Model is well suited for describing and testing hypotheses about relationships between a categorical dependent variable Y and one or more categorical or continuous explanatory variables X. Suppose $\pi_g(x_k) = P(Y = g|x_k)$ is the probability that the gs category of response variable Y(g=1,...,G) for the ks combination of X, the response function is shown as:

$$\pi_g(x_k) = P(Y = g|x_k) = \frac{e^{\beta_g^T x_k}}{\sum_{g=1}^{G} e^{\beta_g^T x_k}}; (g = 1, ..., G) \tag{3.29}$$

In general,for every different two categories r and S, $r \neq s, r, s \neq G$,the Multinomial logit model takes the form that

$$log_e \frac{\pi_r(x_k)}{\pi_s(x_k)} = (\beta_r - \beta_s)^T x_k \tag{3.30}$$

The assumptions are:

- For every combination of X-variable x_k, the response variable Y follows a multinomial distribution with frequency n_k.
- The responses variable's distribution of frequecies for different combinations x_k is independent from one another.

- There exist one simple test sample, that observations in it are independent from one another.

Often when we are doing the regression modelling, we set β_G to zero vector as normalization and thus:

$$\pi_G(x_k) = \frac{1}{\sum_{g=1}^{G} e^{\beta_g^T x_k}}; (g = 1, ..., G) \tag{3.31}$$

As the result, the g logit takes the form:

$$log_e \frac{\pi_g(x_k)}{\pi_G(x_k)} = (\beta_g^T) x_k, (g = 1, ..., G) \tag{3.32}$$

To estimate the coefficient, the maximum likelihood method is widely used. The model and estimation methods are presented in Rönz (2001).

4 Exploratory Factor Analysis of Evaluation Data for Teaching

In this part, we are trying to explore the psychometric properties of students' responses and the degree to which these dimensions may be empirically confirmed using factor analysis.

4.1 Independence

Before making factor analysis of item variables, we use "chi-square test of independence" to identify whether there exist relationships among them or not. We have merged categories with small frequencies in order to ensure that the estimated expected frequency $\hat{e}_{jk}$ of every cell is larger than 1.

The standard of merging is $e = p_{min}^2 * n \geq 1 \Rightarrow p_{min} \geq \sqrt{\frac{1}{n}}$ where p_{min} is the minimum frequency of total item variables. The items we merged for each dataset is shown in Table A.7-A.10.

After the confirmation that the two conditions of "chi-square test" are met, we have calculated the chi-square matrix of item variables. The outcomes have revealed that the hypothesis of variables' independence is largely rejected.

Associations among item variables are also demonstrated by corrected contingency coefficient(CC) and normalized uncertainty coefficient(UC) results, which are presented in attatched CD (directory: appendix/independence).

4.2 Factor Modelling

As students' responses to the the items are not independent from each other, there should be common factors behind the data. This led to the question how many factors are represented by these response items. We have made factor analysis by underlying variable approach. The software we have utilized here is M-plus, which is specially designed for the analysis of categorical data. From outcomes of eigenvalues for sample correlation matrix and cumulative variance they have explained, which are shown in Table A.12 and A.13, and according to the standard factor extraction criterion of an eigenvalue larger than 1, six factors can be extracted from lecture course 2003 sample and five factors from other three datasets. In the mean time, it is worthwhile to note that the first eigenvalue is much larger than others and has explained about 35 percent of the total sample variance. Considering this result we also have selected one-factor model to pursue further analysis. Results under varimax rotation of all models from one-factor model till five-factor are illustrated in attached CD(directory: appendix/factormplus).

4.2.1 Five-factor Model

The solutions for five- or six- factor model in TableA.14 have suggested, first, that although we choose six-factor model for data of lecture course 2003, it is noted that no response item has significant loadings($> 0,5$) on the sixth factor. From Figure 4.1, which shows the loading coefficients results, five common factors underlying responses datasets can be generalized as follows:

Lecturer's communication skill : It consists of items, which are pertaining to teacher's teaching characteristics and ability to teach, e.g., explain ability, content clarity, transparency ability, willingness to answer questions, topic structure clarity and so on.

Quality of lecture notes : It exhibits largely loadings for items relating to the lecture notes which teacher used in the class: quality, choice and availability of the lecture notes.

Course attributes : It is defined by items concerning course attributes, such as speed, mathematical level, difficulty level and challenge level of the course.

Students reactions : It consists of items, which represents the students self assessment, like the interest degree, attention span during the class and knowledge increase.

Question answering : It includes items concerning the instructors' willingness to answer questions and the quality of question answered.

It is understandable that teaching effectiveness has a big influence on the response data. Lecturer's communication skill, quality of lecture notes, question answering and students' reactions are four important fields of instructors' teaching strategy.

	Lecturer								Lecture concept									Attributes			Self assesment					Atmosphere				
	Explain ability	Content clarity	Transparancy quality	Didactical ability	Stimulation of Independent thought	Willingess to answer questions	Quality of answered questions	Time allowed after course&	Aspects covered deepness	Topic structure clarity	Related topics reference	Practical example application	Choice of lecture notes	Availability of lecture notes	Presence in the internet	Content update	Lecture&exercise relevance	Lecture speed	Mathematical level	Difficulty	Interest degree	Attention span	Knowledge increase	Preparation level	Challenging feeling	Atmosphere-Stress level	Atmosphere-interest degree	Atmosphere-disciplined degree	Atmosphere-motivation level	Explained variance
VL 2003	B1	B2	B3	B4	B5	B6	B6_2	B7	C1	C2	C3	C4	C5	C6	C7	C8	C9	D1	D2	D3	E1	E2	E3	E4	E5	F1	F2	F3	F4	
Communication skill	,7	,7	,6	,6	,4	,3	,4	,1	,7	,7	,5	,4	,4	,2	,2	,4	,0	,0	,0	-,1	,2	,2	,4	-,1	,0	,2	,4	,3	,5	0,17
Lecture notes	,1	,2	,3	,1	,1	,1	,1	,2	,3	,3	,2	,3	,6	,8	,7	,4	,1	,0	,0	,0	,1	,1	,2	-,1	,0	,1	,0	,1	,0	0,08
Course attributes	,1	,2	,0	,0	,0	,0	,0	,0	,0	,0	,0	,1	,0	,0	,0	,0	,1	-,7	-,7	-,9	,1	-,1	,0	,3	-,6	,3	,1	,1	,1	0,09
Question answering	,4	,3	,2	,4	,3	,8	,8	,2	,1	,1	,2	,3	,1	,2	,2	,3	,3	,0	,0	,0	,1	,2	,2	,1	,0	,2	,3	,3	,3	0,09
Students reactions	,3	,3	,1	,3	,4	,2	,2	,0	,3	,2	,3	,3	,2	,1	,1	,4	,1	,0	-,1	-,1	,7	,7	,7	-,2	,0	,2	,7	,3	,6	0,13
VL 2002																														
Communication skill	,8	,7	,5	,7	,4	,2	,4	,1	,6	,7	,5	,4	,3	,1	,2	,3	,1	,0	-,1	-,1	,1	,2	,4	,1	,0	,2	,5	,3	,5	0,16
Lecture notes	,1	,2	,4	,1	,1	,1	,2	,1	,3	,4	,3	,4	,7	,8	,7	,5	,1	,0	-,1	-,1	,1	,1	,2	-,1	,0	,1	,1	,2	,2	0,10
Course attributes	,2	,3	,1	,1	,0	,1	,1	,0	,0	,1	,1	,2	,1	,1	,0	,1	,3	-,8	-,8	-,9	,1	-,1	,1	,3	-,5	,4	,1	,1	,2	0,10
Question answering	,3	,2	,1	,4	,4	,8	,8	,3	,1	,1	,2	,3	,1	,2	,1	,3	,3	-,1	,0	,0	,1	,2	,2	,1	,0	,2	,2	,2	,3	0,08
Students reactions	,3	,2	,2	,3	,4	,2	,2	,0	,2	,2	,3	,3	,2	,1	,1	,3	,1	,0	-,1	-,1	,7	,7	,7	-,3	,0	,2	,7	,4	,6	0,12
UE 2003																														
Communication skill	,8	,7	,6	,7	,5	,7	,8	,2	,4	,4		,3	,2	,1	,2			,0	,0	,0		,1	,2	,0	,0	,2	,2	,2	,3	0,17
Course attributes	,0	,2	,0	-,1	-,1	-,1	-,1	,1	,1	,1		,1	,1	,0	,0			-,8	-,7	-,9		-,3	-,1	,3	-,7	,3	,0	,1	,1	0,11
Lecture notes	,1	,1	,2	,1	,1	,2	,2	,2	,3	,3		,2	,6	,9	,7			,0	,0	,0		,1	,1	,0	,0	,1	,1	,2	,1	0,09
Concept	,3	,4	,3	,3	,2	,0	,0	,0	,7	,7		,4	,3	,1	,1			,0	-,1	-,1		,4	,5	-,2	,0	-,1	,3	,0	,3	0,09
Students reactions	,2	,2	,1	,2	,4	,3	,3	,1	,2	,2		,3	,1	,1	,1			,0	-,1	-,1		,5	,6	-,1	,0	,5	,7	,6	,7	0,11
UE 2002																														
Communication skill	,8	,6	,5	,7	,5	,7	,8	,3	,3	,4		,4	,2	,2	,2			-,1	,0	,0		,2	,3	,0	,0	,2	,2	,3	,3	0,17
Course attributes	,1	,2	,1	,0	-,1	,0	,0	,0	,1	,1		,2	,2	,1	,0			-,8	-,7	-,9		-,3	,0	,3	-,7	,3	,1	,0	,2	0,12
Lecture notes	,0	,1	,3	,1	,1	,2	,2	,2	,2	,3		,2	,7	,9	,7			,0	-,1	-,1		,0	,1	,0	-,1	,2	,1	,2	,2	0,10
Students reactions	,3	,3	,2	,3	,4	,3	,2	,0	,3	,2		,3	,2	,2	,2			-,1	-,1	-,1		,6	,6	-,2	,0	,5	,8	,5	,8	0,14
Concept	,3	,4	,3	,3	,2	,0	,0	,1	,7	,7		,3	,2	,1	,1			-,1	-,1	-,1		,2	,3	-,1	,0	,0	,2	,0	,1	0,07

Figure 4.1: Loadings of five-factor model, red: loading $\in [0,7,1]$, blue: loading $\in [0,5,0,7[)$.

It is worthwhile to note that except teaching quality, course attributes to the class do as well affect the students' response patterns to some extent. In this area, it is very hard for instructors to affect the ratings they received through the improvement of teaching quality.

Contrast to the items which are highly correlated with the factors, it is noted that several items do not highly load on any factors, e.g., stimulation of independent thought, time allowed after course, content update, relevance between lecture and exercise, preparation level of students and stress level of class atmosphere.

One possible reason for insignificance of some item variables, e.g., "time allowed after course" and "relevance between lecture and exercise", is that, during imputation process, we have created new category "0" for missing values, making the response data for these variables not ordered anymore. This will obviously lead error to the result of analysis. Another explanation of the irrelevance, is that, though these items do have some influence on the students' ratings in fact, they are not so significant aspects which students care about when they give evaluation scorings. In another word, they are not significant fields that instructors should pay much attention to make an effective teaching process.

4.2.2 One-factor Model

From the Figure 4.2, which has illustrated the loadings result of one-factor model, it is obvious to see that items concerning the course attributes and students attitude have very small discrimination coefficient, indicating they are not highly correlated with the factor. Meanwhile, the rest items related to some very important aspects of

	Chi-square value	Degree of freedom	Critical value (95%)
One-factor-model	16576	160	191
Two-factor-model	10453	165	196
Three-factor-model	7079	182	214
Four-factor-model	4894	186	219
Five-factor-model	3519	177	209
Six-factor-model	2334	169	200

Table 4.1: Chi-square test of model fitting

teaching: the teacher's ability to teach, quality and availability of lecture notes and course atmosphere have shown strong associations with the latent variables. By this distribution feature we can say that the single factor represent the general ability of lecturer to teach. Taking this assumption, we can pursue further discussions about the determination of ranking of teaching effectiveness in the next sections.

4.2.3 Model Fitting

Judging by the large chi-squared residuals observed in the two-way margins, the factor models of all four datasets are surprisingly meeting with bad fits. The result of chi-square test value of models for lecture course 2003 data is shown in Table 4.1 as an example.

There are a number of possible reasons why the factor model for ordinal responses is not a good fit and the facts are given below.

- *Imputation process.* When we deal with the missing value imputation, we recode missing value of items which has high percentage into 0, and that makes the categories not ordered anymore.
- *Response pattens.* When the number of variables is large, many response patterns will have expected frequencies which are very small. This will make the condition of chi-squared test that:

UE 2002	UE 2003	VL 2002	VL 2003		Item	Group
,8	,8	,8	,8	B1	Explain ability	Lecturer
,8	,8	,8	,8	B2	Content clarity	
,6	,6	,7	,6	B3	Transparancy quality	
,8	,8	,8	,8	B4	Didactical ability	
,6	,6	,6	,7	B5	Stimulation of Independent thought	
,6	,7	,7	,7	B6	Willingess to answer questions	
,7	,8	,8	,8	B6_2	Quality of answered questions	
,2	,2	,2	,2	B7	Time allowed after course&	
,7	,7	,7	,7	C1	Aspects covered deepness	Lecture concept
,7	,8	,7	,8	C2	Topic structure clarity	
		,6	,7	C3	Related topics reference	
,6	,6	,7	,7	C4	Practical example application	
,7	,6	,6	,6	C5	Choice of lecture notes	
,7	,6	,6	,6	C6	Availability of lecture notes	
,6	,5	,6	,6	C7	Presence in the internet	
		,6	,7	C8	Content update	
		,4	,3	C9	Lecture&exercise relevance	
-,4	-,2	-,3	-,1	D1	Lecture speed	Attributes
-,5	-,3	-,4	-,2	D2	Mathematical level	
-,6	-,3	-,5	-,3	D3	Difficulty	
		,5	,6	E1	Interest degree	Self assesment
,4	,5	,6	,6	E2	Attention span	
,7	,7	,7	,8	E3	Knowledge increase	
,0	-,1	,0	-,1	E4	Preparation level	
-,4	-,2	-,2	-,1	E5	Challenging feeling	
,5	,4	,5	,4	F1	Atmosphere-Stress level	Atmosphere
,7	,7	,8	,8	F2	Atmosphere-interest degree	
,5	,5	,6	,5	F3	Atmosphere-disciplined degree	
,8	,8	,8	,8	F4	Atmosphere-motivation level	
0,39	0,35	0,38	0,38		Explained variance	

Figure 4.2: Loadings of one-factor model, red: loading $\in [0,7,1]$, blue: loading $\in [0,5,0,7[)$.

"estimated frequency should be larger than 1 " not satisfied. Test statistics will not follow the chi-squared distribution any more and from practical point of view these tests cannot be carried out.

5 Measurement of Teaching Effectiveness

With the purpose to identify the ranking of the lecturer according to their teaching effectiveness, a general standard criterion generated from evaluation data is necessary.

In one-factor model analysis, the single factor is closely reflecting instructors' general teaching ability. It then turns out the idea to make use of this factor as the single indicator of the lecture's teaching effectiveness. Based upon the mean value of factor score of each course, the ranking of teaching effectiveness of each instructor can be determined.

5.1 Score Calculation

Unfortunately, we have no software in hand to calculate the factor score of categorical data. One way to solve this problem is to use SPSS instead, treating the data as continuous. First we make one-factor model analysis in SPSS, and then compare the loadings results with that got from M-plus. The correlation coefficients between two loading results, listed in Table 5.1, have revealed high correlations between them. According to this consequence, we assume fac-

	Correlation	Significance level
Lecture course 2003	0,997	0,00
lecture course 2002	0,997	0,00
Exercise course 2003	0,992	0,00
Exercise course 2002	0,994	0,00

Table 5.1: Correlations of loading results between Mplus and SPSS.

tor scores calculated in SPSS, treating the data continuous, coincide with the factor scores calculated when data are treated as categorical. The detailed results of one-factor model analysis using SPSS and loading comparisons are shown in attached CD (directory: appendix/factorspss/onefactor).

5.2 The Lectures' Rank of Teaching Effectiveness

Depending on the mean value of factor scores, the ranking of teaching effectiveness is determined. The larger the score is, the lower rank the course has received. Detailed rank results of courses of four datasets are displayed in attached CD (directory: appendix/factorspss/rank).

It is obviously observed from Table 5.2, which shows the list of the best and worst five lectures courses in 2003, that all courses in best group are in graduate level with small class size, e.g., the number of students in the first two best courses(”aaa”, ”bbb”) is less than 10. One the contrary, the first worst and second worst course(”fff” and ”ggg”) are in undergraduate level with over hundreds students in the class. It is also worthwhile to note that professors also have made bad

	Course code	score	std.dev.	class size	level	time	teacher
1	aaa	-1,66	0,36	9	graduate	afternoon	Prof.
2	bbb	-1,45	0,35	8	graduate	morning	Prof.
3	ccc	-0,97	0,46	47	graduate	afternoon	Prof.
4	ddd	-0,89	0,58	40	graduate	afternoon	Assist.
5	eee	-0,88	0,79	11	graduate	morning	Prof.
	Course number	score	std.dev.	class size	level	time	teacher
1	fff	1,24	0,87	113	undergrad.	morning	Prof.
2	ggg	0,90	1,05	247	undergrad.	morning	Prof.
3	hhh	0,73	0,99	50	graduate	morning	Prof.
4	iii	0,64	0,93	95	graduate	afternoon	Assist.
5	jjj	0,52	0,93	78	graduate	afternoon	Prof.

Table 5.2: The best (above) and worst(bottom) 5 lectures in 2003.

courses. In some sense, it represents high level of knowledge is not sufficient for man to be a good instructor.

5.2.1 Score and Course Attributes

From teaching effectiveness ranking results above, one interesting topic is led to further study: What kind of courses is more attractive to students? Still taking the datasets sample for lecture course 2003, we have compared the differences of the score courses received with their different own characteristics.

Class Level and Class Size

As portrayed in Figure 5.1 and Figure 5.2, the mean score of courses in undergraduate is much lower than in graduate level. Meanwhile, courses with small size have received relatively higher score. Compared to compulsory courses, optional courses are more welcomed by students, which is confirmed by Figure 5.3,

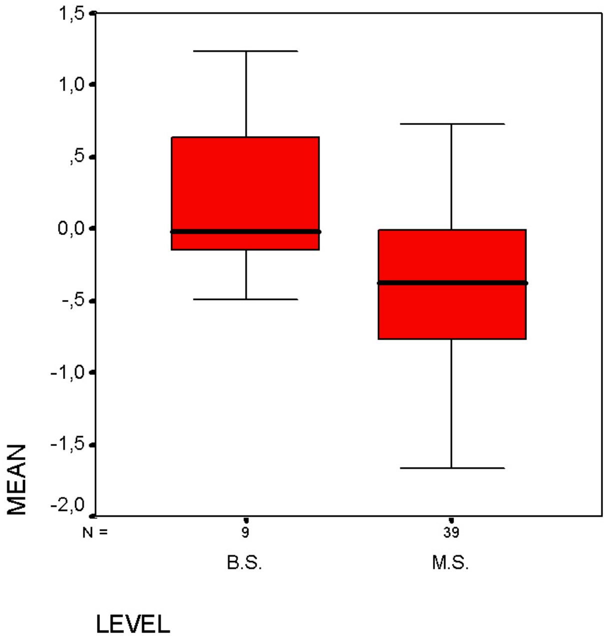


Figure 5.1: Mean score vs. level

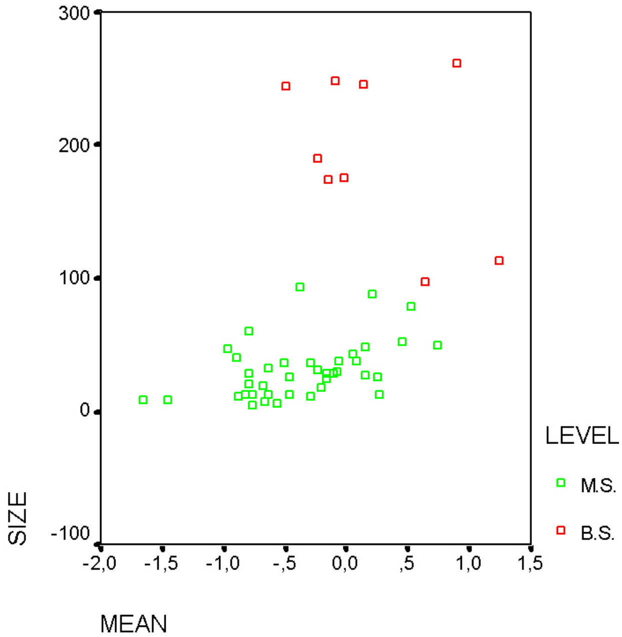


Figure 5.2: Mean score vs. class size.

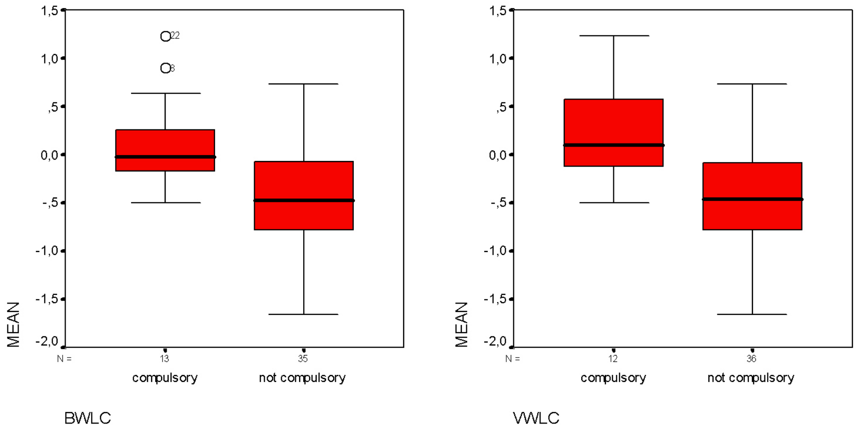


Figure 5.3: Mean score vs. compulsory.

Class Meeting Time

Figure 5.4 depicts the distribution of mean scores of courses with respect to time they meet. It can be observed that the afternoon class and class which are arranged in the middle of week have received higher score than others.

Instructor

Figure 5.5 reveals that, on average, assistant teacher get higher score than professors. Although the number of assistant lecturer in our data sample are much more less than professors, which will lead some bias to our results, we still can get some ideas that the rank of the instructor does not matter a lot to the evaluation responses of students.

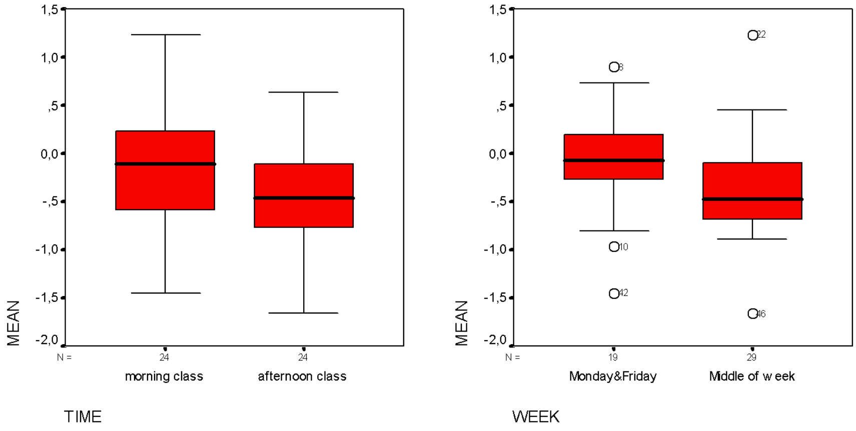


Figure 5.4: Mean score vs. class meeting time.

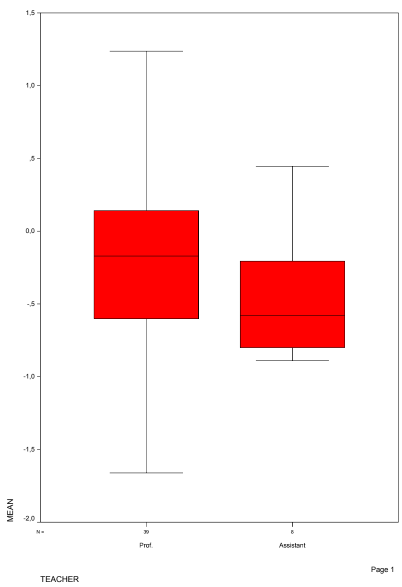


Figure 5.5: mean score vs. teacher.

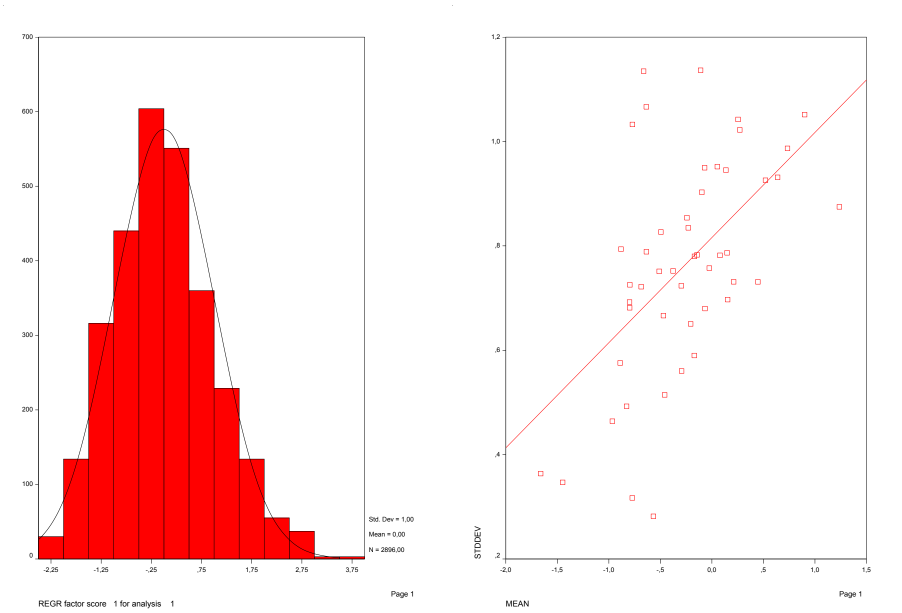


Figure 5.6: F. score distribution(left), relation of mean and stddev(right).

5.3 Score Distribution

When we are looking at the distribution of factor score results of all courses, it is unexpectedly that it does not follow the normal distribution, see Figure 5.6. The possible reason for that is the response datasets are positively skewed.

According to the ranking results we have picked out two course sample from lecture course 2003 dataset, one with relatively low score ("fff") and another with high score of teaching effectiveness("ccc"), see in Table 5.2. Looking at Figure 5.7, which has illustrated the distribution of factor score across the students. It is clear to note that the variance of bad courses are much higher than good course, the scores ranges from bad to good level. On the other hand, for good course, scores are more concentrated on good level.

From the kernel density approximation of the factor score of both

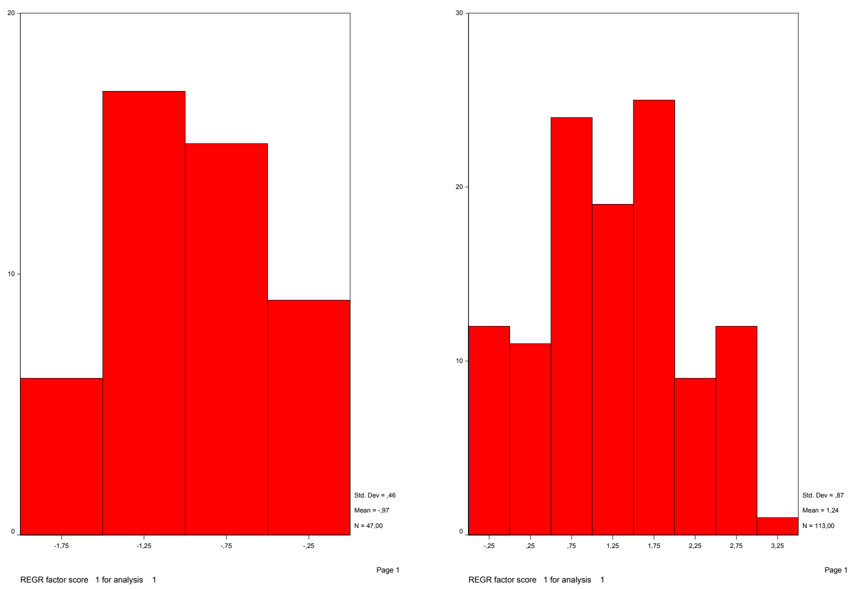


Figure 5.7: Fator score, left: good course, right: bad course .

course, choosing the different bandwidth, which is depicted the in Figure 5.8. It is confirmed that students' response to bad course distributed not smoothly as to good course. Several response modes have appeared in the whole range. On the contrary, only one mode comes into sight among the response data for good course. From this distribution character, we can suggest, in one class of bad course, students can be divided into several groups depending on their different ratings to the course. For the reasons of time limitation, we do not make further discussions of this problem. But it is worthy to give more study in this area.

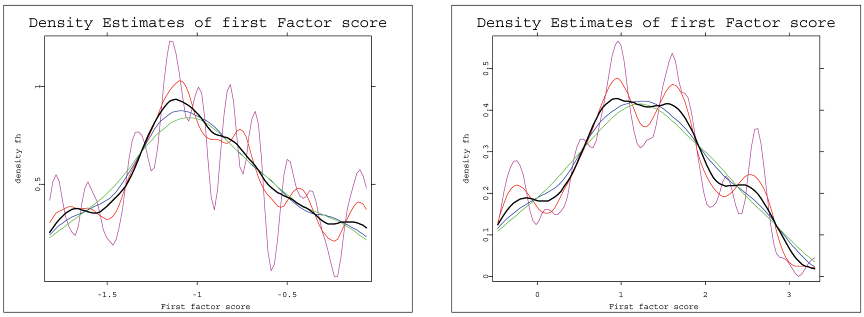


Figure 5.8: Density estimation of F.Score, left:good course, right:bad course.

kernel.xpl

5.4 Method Discussion

To use this measure, we have to pay much attention to some problems. First, Teachers get different evaluation score in different years, and the loading of the model will be also changed when we make factor analysis. This significant inconsistency make it difficult for us to identify whether the teaching effectiveness of one course is improved or not. In this case, the loadings of one-factor model is necessary to be averaged to achieve independence of time scale. Second, there still exist a lot of problems in this method, such as the model's misfitting, treating the data as continuous in factor score calculating. The reliability of this method still requires further examination.

Our finding also sheds light on possible sources of student evaluations. Still by factor analysis on SET, we recognize five separate dimensions of instructional effectiveness, namely lecturer's communication skill, quality of lecture notes, course attributes, students reactions and question answering. But as a complex multidimensional activity, teaching also comprises of a number of a separable variables such as teacher's characteristics (e.g., gender, reputation, teaching experience), course characteristics (e.g., meeting time, difficulty level, class size) and students' characteristic (e.g., gender, major). SET instruments should also reflect this multidimensionality. By employing multinomial logic regression techniques, we find that it measures not only aspects of instructional effectiveness, but also captures some factors that are normatively considered irrelevant to teaching quality.

6 Determinants of SET

We begin our analysis by asking what factors affect the evaluation scores. Particularly, we want to examine whether factors that are normatively irrelevant to teaching quality matter or not. The model we use in this study is multinomial logistic model.

The dependent variable is item variable "Global overall ratings of the course", which scaled from 0 to 4, and it takes on 0 if the response is missing.

The explanatory variables included in the model are the following:

- *The factor scores of three-factor model* indicating the teaching performance, they are continuous.

  Here we meet with the same problem that no software in hand to calculate the factor scores of categorical data. We are still using SPSS instead, treating the data as continuous. After comparing its loadings result with that got from M-plus, from the correlation coefficient between two loading result, listed in Table A.15, we have found that not all the factors are highly correlated. To solve this problem, we use three-factor model because all factors calculated by two methods are highly correlated (the correlation coefficient of loadings are larger than $0,95$), the result of correlation coefficient are shown in Table A.16. The main factors are implement in the model are:

- Teaching ability and students reactions
- Lecture notes.
- Course attributes

The loading results for all datasets are listed in Figure 6.1. Results of five-factor and three-factor model analysis using SPSS and loading comparisons are shown in attached CD(directory: appendix/factorspss/morefactor).

- *Class size:* We standardize its continuous value in to range $[0, 1]$ by

$$N_{std} = \frac{n}{n_{max}} \tag{6.1}$$

where n is the number of students in the class of each data sample.

- *Student major,* taking on a value of 1 if the major is management, 2 economics, 6 otherwise.
- *Student gender,* taking on a value of 1 if the student is female, 2 otherwise.
- *Class time,* taking on a value of 2 if the class meets before 2:00pm, 3 after 2:00pm.
- *Day of class,* taking on a value of 1 if the class meets on Monday or Friday, 2 otherwise.
- *Class level,* taking on a value of 1 if it is undergraduate class, 2 graduate class.
- *Class compulsory for economic students,* taking on a value of 1 if it is compulsory, 2 otherwise.
- *Class compulsory for management students,* taking on a value of 1 if it is compulsory, 2 otherwise.

	Lecturer								Lecture concept									Attributes			Self assesment					Atmosphere				
	Explain ability	Content clarity	Transparancy quality	Didactical ability	Stimulation of Independent thought	Willingess to answer questions	Quality of answered questions	Time allowed after course&	Aspects covered deepness	Topic structure clarity	Related topics reference	Practical example application	Choice of lecture notes	Availability of lecture notes	Presence in the internet	Content update	Lecture&exercise relevance	Lecture speed	Mathematical level	Difficulty	Interest degree	Attention span	Knowledge increase	Preparation level	Challenging feeling	Atmosphere-Stress level	Atmosphere-interest degree	Atmosphere-disciplined degree	Atmosphere-motivation level	Explained variance
VL 2003	B1	B2	B3	B4	B5	B6	B6_2	B7	C1	C2	C3	C4	C5	C6	C7	C8	C9	D1	D2	D3	E1	E2	E3	E4	E5	F1	F2	F3	F4	
Communications&stud.reaction	,6	,6	,4	,6	,6	,5	,6	,0	,5	,5	,5	,5	,3	,0	,1	,5	,2	,0	-,1	-,1	,7	,7	,8	-,2	,0	,3	,8	,4	,8	0,24
Lecture notes	,4	,5	,6	,4	,2	,3	,4	,3	,5	,5	,4	,4	,7	,8	,8	,5	,2	,0	,0	-,1	-,1	,0	,2	,0	,0	,2	,1	,2	,2	0,15
Course attributes	-,1	-,2	,0	-,1	,0	,0	,0	,0	-,1	-,1	-,1	-,2	,0	,0	,0	,0	-,1	,8	,7	,8	,0	,1	,0	-,4	,6	-,4	-,1	,0	-,1	0,09
VL 2002																														
Communications&stud.reaction	,7	,6	,4	,7	,7	,5	,6	,1	,5	,5	,5	,4	,2	,0	,1	,4	,2	,0	-,1	-,2	,6	,7	,7	-,1	,0	,3	,8	,5	,8	0,23
Lecture notes	,3	,4	,6	,4	,2	,3	,3	,2	,5	,5	,4	,5	,8	,8	,8	,6	,2	,0	-,1	-,1	,0	,0	,2	,0	,0	,2	,2	,3	,3	0,15
Course attributes	,2	,3	,1	,2	,0	,1	,2	,0	,1	,1	,1	,2	,1	,1	,0	,0	,4	-,8	-,7	-,8	,0	-,2	,1	,4	-,5	,5	,1	,1	,2	0,11
UE 2003																														
Communications&stud.reaction	,7	,7	,5	,7	,6	,6	,6	,1	,6	,6		,5	,2	,1	,1			,0	-,1	,0		,6	,7	-,2	,0	,4	,7	,5	,8	0,22
Lecture notes	,2	,3	,4	,3	,1	,3	,3	,3	,4	,4		,3	,7	,8	,8			-,1	,0	-,1		,0	,1	,0	-,1	,0	,0	,0	,1	0,10
Course attributes	,0	-,1	,1	,1	,2	,1	,1	,0	-,1	-,1		-,1	-,1	,0	,1			,8	,7	,9		,3	,0	-,3	,6	-,4	-,1	-,1	-,2	0,10
UE 2002																														
Communications&stud.reaction	,8	,7	,6	,7	,7	,6	,6	,1	,5	,5		,6	,3	,2	,2			-,1	-,1	-,1		,6	,7	-,2	,0	,5	,7	,5	,7	0,26
Course attributes	,0	-,2	-,1	,0	,1	,1	,1	,1	-,2	-,2		-,2	-,2	-,1	,0			,8	,8	,9		,3	-,1	-,4	,7	-,4	-,1	,0	-,2	0,12
Lecture notes	,1	,2	,4	,2	,1	,2	,2	,3	,3	,4		,3	,8	,9	,8			,0	,0	-,1		,0	,1	,1	,0	,1	,1	,1	,1	0,11

Figure 6.1: Loadings of three-factor model, red: loading $\in [0,7,1]$, blue: loading $\in [0,5,0,7]$).

- *Instructor's rank,* taking on a value of 1 the teacher is professor, 2 assistant.

The model estimates of four datasets are presented in Figure 6.2. The complex structure of the model makes it hard to judge how large an affect a variable has on the scores from simply looking at the coefficient estimates. However, we can still identify which variables have significant effects from the significance of the Wald statistic (e.g., less than 0.05). Here we take the dataset of lecture course 2003 for example to discuss in detail.

It is obvious to see that the communication skill of instructors and students reaction of teaching has biggest influence to evaluation scores. The core quality that good teachers possess is the ability to communicate their knowledge and expertise to their students. Effective teaching activities, such like making the class interesting and more receptive, trying the best to attract student's attention and letting students feel their knowledge increase a lot should have positive effect on student ratings.

The estimated coefficients indicate that quality of lecture notes are second important aspects of teaching effectiveness. As one communication instrument of teaching, lecture notes play a very import role in teaching performance. Clear and organized notes will help students understand materials. Putting more effort in the preparation of lecture notes is an efficient way to improve the teaching effectiveness.

Except teaching performance, the course own attributes, e.g., mathematical level, difficulty level , the meeting time of course also have significant effect on student ratings. The people who teach in afternoon and choose interesting courses without too much maths teach

		Observations	Intercept	Communication&students reaction	Course attributes	Lecture notes	Class size	BWL- others	VWL – others	Female	BWL compulsory	VWL compulsory	B.S.	morning class/afternoon class	Monday or Friday	Professor
VL 2003		2896														
	0	268	2,37	-3,20	-1,76	-0,55	1,40	-0,83	-0,51	-0,23	0,59	0,01	-0,64	-0,62	0,00	0,39
	1	398	0,16	-5,99	-3,42	-1,05	1,36	0,03	0,84	0,14	0,37	1,03	-0,50	-0,83	-0,10	-0,42
	2	1248	3,36	-3,62	-2,02	-0,64	0,77	-0,11	0,37	0,20	0,73	0,88	0,80	-0,60	-0,21	0,01
	3	703	2,76	-1,65	-0,96	-0,27	0,27	-0,31	0,13	-0,05	0,54	0,44	-0,70	-0,41	-0,04	0,41
	4	279														
VL 2002		3247														
	0	347	1,70	-2,54	-1,41	-0,53	-0,11	-0,39	-0,07	-0,12	-0,17	1,13	-0,78	0,53	-0,50	-0,43
	1	265	-1,90	-5,59	-3,35	-1,49	2,21	-0,52	-0,09	0,31	-0,47	0,38	0,57	0,60	-0,45	-0,65
	2	1314	1,06	-3,52	-2,10	-0,97	1,91	-0,07	0,15	0,13	0,09	0,72	-0,15	0,72	-0,31	-0,06
	3	844	1,99	-1,70	-0,98	-0,48	0,66	-0,44	-0,11	-0,18	0,26	0,50	-0,37	0,60	-0,24	-0,12
	4	477														
UE 2003		1980														
	0	181	1,65	-3,33	-1,22	-0,05	0,41	0,41	0,54	-0,56	0,29	0,64	0,32	0,38	-0,45	0,17
	1	330	-0,94	-5,57	-2,17	0,00	0,73	0,70	0,72	0,18	0,24	0,49	1,12	0,94	0,12	0,02
	2	982	2,82	-3,34	-1,21	-0,12	0,71	0,47	0,54	-0,03	0,20	0,64	0,48	0,39	0,15	0,01
	3	369	1,97	-1,40	-0,44	-0,18	0,41	0,32	0,69	-0,12	0,32	-0,29	1,02	-0,34	0,11	0,19
	4	118														
UE 2002		2455														
	0	286	2,82	-2,49	-0,26	-0,42	0,49	-0,22	-0,39	0,01	-0,90	-0,05	0,66	0,30	0,18	-0,28
	1	273	-0,58	-5,86	-1,04	-1,42	0,84	-0,17	0,07	-0,24	-1,06	1,35	0,68	-0,04	0,38	-0,38
	2	1237	2,87	-3,30	-0,63	-0,68	0,80	0,14	-0,18	-0,08	-0,83	0,57	0,96	0,19	0,50	-0,15
	3	502	3,05	-1,45	-0,27	-0,24	0,73	-0,08	-0,30	-0,10	-0,80	-0,14	0,81	-0,27	0,23	-0,24
	4	127														

Figure 6.2: Coefficients of Multi. Logit model, red: sig.$< 0,01$, blue: $0,01 <$sig.$< 0,05$

will be more possible to receive high ratings. Could it be that students willing to take late and relatively easy classes more receptive to the efforts of their instructors?

Although from modelling estimates of dataset lecture course 2003, the size of the class is not significant, but results of other three datasets have shown that the courses with small number of students are more willing to receive higher scores to some extent. This disagreement results in the imperfection of our datasets that sample is too small. Whether class size does influence the evaluation score or not need further research.

The estimation result also show that other explanatory variables which are unrelated to the teaching quality, such as week time of the class, compulsory or not, rank of teacher, major and gender of students have no bearing on the students' overall ratings at all.

Detailed results of Multinomial logit model are displayed in attached CD (directory: appendix/model)

7 Conclusion

In this paper, we propose to apply two advanced statistical techniques, namely *exploratory factor analysis* of categorical data and *multinomial logit model*, on the student evaluation data, to assess the effectiveness of teaching at higher education institutions in a quantitative approach.

In the one-factor model, where the factor represents the general teaching ability of instructors, a single numerical scoring device is created to evaluate the teaching performance of the lecturers. The ranking of the teaching effectiveness of each lecture, which is generated by this method, reveals that the courses that meet the following requirements have higher probability to receive high scores.

1. Small class size
2. Offered at graduate level
3. Optional to students
4. Meet at the afternoon
5. Meet in the middle of the week

A close examination on the evaluation data of the courses with relatively low score and ones with high score further discloses that the variance of the scores for the low score course is much larger than that of the high score course, which in essense implies that students' opinions on a possibly poorly deliverd course tend to differ more significantly than those on a possibly well taught course.

Based upon the empirical factor analysis of student survey data, this paper has identified following five main determinants, which can significantly affect SET scores.

1. Lecturer's communication skill
2. Quality of lecture notes
3. Course attributes
4. Students' reactions
5. Question answering

The results generated by the multinomial logic regression shows that

1. Communication skill & students' reaction
2. Course attributes
3. Quality of lecture notes

are three most important factors which determine the student response to "general overall ratings" of the course. Meanwhile, class meeting time and class size, which are normatively considered irrelevant to teaching quality, may also have perceivable effect on the ratings.

Even though the initial targeted application of this quantitative approach is the assessment of teaching effectiveness in higher education institutions, it is worthwhile to note that the fundamental methodology can also be extended to evaluate the quality of education in primary schools, high schools and vocational schools.

There are several minor issues remain to be resolved in this study. First the data used neither represent random sample and nor is it complete (e.g., data of some lecture and exercise courses are missing). Second, during the imputation process, the creation of new category "0" for missing values to some item variables will make the data not ordinal anymore, and thus may introduce potential error into the analysis. Third, the factor model we have built is far away from fitting. Last but not the least, the continuity of the data has been assumed when the factor scores are calculated. The reliability of this study may require further investigation and verification.

A Appendix

One CD containing datasets, XploRe programms compiled and important results of analysis is attached to the dissertation. We list all the table which we have referred to in the text in this chapter.

- Table A.1 - A.5 give out the list of tables which describe the characteristics of the data set we use.
- Table A.6 shows the percentage of missing values in each item variable.
- Table A.7 - A.10 are the frequency tables which show the students response patterns for all four datasets. And frequency values are in percentage.
- Table A.11 shows the variable code used in the analysis.
- Table A.12 and Table A.13 have listed the eigenvalues result of factor analysis and the variance of data they explained.
- Table A.14 show the factor structures in five-factor model: the value x in the table represent the xth factors in five-factor model, whose loadings is larger than $0,5$ and smaller than $0,7$, and x^* means the loading on Xth factor is larger than $0,7$.
- Table A.15 and Table A.16 have shown the correlation coefficients between factor loadings calculated from SPSS and M-plus for five-factor model and three-factor model.

	Minimum	Maximum
Lecture2002	4	264
Lecture2003	4	262
Exercise2002	9	270
Exercise2003	5	329

Table A.1: Overview of Class size.

	Instructor's gender		Instructor's Rank	
	Male	Female	Professor	Assistant
Lecture2002	89,9	10,1	88,5	11,5
Lecture2003	90,8	9,2	85,3	14,7
Exercise2002	74,0	26,0	1,1	98,9
Exercise2003	87,5	12,5	0,2	99,8

Table A.2: Percentage of instructors' character variables.

	Number of	Major			Gender	
	Observations	Management	Economics	others	male	female
Lecture2002	3247	53,1	28,3	18,6	49,6	50,4
Exercise2002	2455	50,1	30,8	19,6	52,8	47,2
Lecture2003	2897	55,1	25,5	19,34	50,3	49,7
Exercise2003	1980	54,7	29,3	15,9	49,1	50,9

Table A.3: Percentage of students' character variables.

	Compulsory		Class level	
	Management	Economics	Undergraduate	Graduate
Lecture2002	63,9	64,9	57,8	42,2
Lecture2003	63,1	66,5	58,5	41,5
Exercise2002	81,7	90,5	82,7	17,3
Exercise2003	77,4	81,2	70,2	29,8

Table A.4: Percentage of course level variables.

	Class time		Class week time	
	Morning class	afternoon class	Monday&Friday	Middle of week
Lecture2002	56,2	43,8	31,5	69,5
Lecture2003	64,5	35,5	49,1	50,9
Exercise2002	79,1	20,9	33,2	66,8
Exercise2003	71,2	28,8	24,4	75,6

Table A.5: Percentage of course time variables.

	Lecture		Exercise	
	2002	2003	2002	2003
Major	1,8	1,7	1,4	0,5
Sex	5,7	4,2	3,9	3,1
Global overall ratings of the course	11,2	9,8	12,1	9,3
Course missing times	3,4	3,0	2,7	2,3
Course missing reason	41,0	42,7	55,3	54,1
Explain ability	0,6	0,5	0,5	0,3
Content clarity	0,6	0,5	0,6	0,6
Transparancy quality	1,0	0,9	1,0	1,0
Didactical ability	1,2	1,4	1,4	2,3
Stimulation of independent thought	1,3	0,8	1,5	1,2
Willingess to answer questions	3,0	3,1	1,3	2,5
Quality of answered questions	5,1	4,8	4,3	4,2
Time allowed after course	23,2	29,7	28,6	28,5
Aspects covered deepness	2,3	2,1	2,2	1,9
Topic structure clarity	1,3	1,2	1,7	1,4
Related topics reference	5,1	6,4	-	-
Practical example application	1,3	1,4	2,7	2,3
Choice of lecture notes	2,3	2,6	4,3	5,2
Availability of lecture notes	4,0	3,6	6,9	6,9
Presence in the internet	3,9	2,9	6,0	5,2
Content update	5,3	5,5	-	-
Relevance beween lecture and exercise	33,5	29,1	-	-
Lecture speed	1,6	1,8	0,7	0,6
Mathematical level	2,3	2,5	1,3	1,4
Difficulty	1,8	1,7	1,3	0,7
Interest degree	0,9	0,9	-	-
Attention span	0,9	0,9	0,9	0,3
Knowledge increase	1,1	1,5	1,6	0,8
Preparation level	6,2	5,9	5,3	4,6
Challenging feeling	14,1	13,2	8,7	8,8
Atmosphere-Stress level	1,7	1,7	1,3	0,6
Atmosphere-interest degree	2,0	1,7	1,3	0,6
Atmosphere-disciplined degree	1,8	1,7	1,6	0,6
Atmosphere-motivation level	1,9	1,7	1,5	0,7

Table A.6: Missing value percentage of data set.

	0	1	2	3	4	5	Mode	Entropy	Merge
Global overall ratings of the course	9,3	13,7	43,1	24,3	8,2	1,4	2	0,18	4/5
Course missing times		38,9	29,4	18,5	8,6	4,5	1	0,23	
Course missing reason	42,2	4,8	6,7	4,5	7,3	34,4	0	0,24	
Explain ability		21,5	46,5	22,3	8,1	1,7	2	0,26	4/5
Content clarity		16,6	46,3	25,3	9,6	2,2	2	0,24	
Transparancy quality		19,3	40,7	26,5	10,8	2,7	2	0,24	
Didactical ability		17,3	42,0	28,7	9,6	2,4	2	0,24	
Stimulation of independent thought		15,1	36,6	34,4	11,2	2,7	2	0,24	
Willingess to answer questions		30,7	48,5	16,8	3,2	0,8	2	0,22	4/5
Quality of answered questions		22,3	49,9	22,0	4,7	1,2	2	0,21	4/5
Time allowed after course	28,8	9,1	26,6	30,3	4,2	1,0	3	0,25	4/5
Aspects covered deepness		14,4	47,3	28,4	8,1	1,8	2	0,23	4/5
Topic structure clarity		18,1	41,9	26,3	10,1	3,6	2	0,24	
Related topics reference		11,7	42,2	34,8	8,5	2,8	2	0,23	
Practical example application		22,7	40,6	25,0	9,1	2,6	2	0,24	
Choice of lecture notes		18,6	39,5	25,7	12,6	3,6	2	0,25	
Availability of lecture notes		24,8	40,8	22,3	8,8	3,3	2	0,24	
Presence in the internet		27,0	41,8	21,0	7,4	2,8	2	0,24	
Content update		19,4	47,6	27,7	4,3	1,0	2	0,23	4/5
Relevance beween lecture and exercise	28,3	18,5	30,1	16,7	4,8	1,6	2	0,26	4/5
Lecture speed		5,6	26,9	59,3	7,1	1,1	3	0,21	4/5
Mathematical level		7,9	25,6	59,2	6,5	0,9	3	0,20	4/5
Difficulty		5,7	30,6	58,5	4,7	0,5	3	0,18	4/5
Interest degree		23,7	38,1	24,4	10,0	3,7	2	0,23	
Attention span		22,3	41,9	24,1	9,4	2,4	2	0,24	
Knowledge increase		12,3	38,8	33,3	11,6	4,0	2	0,24	
Preparation level		16,8	43,6	21,7	9,8	8,5	2	0,25	
Challenging feeling	12,4	4,7	23,1	53,8	5,4	0,6	3	0,23	4/5
Atmosphere-stress level		21,2	33,4	28,8	14,1	2,4	2	0,24	
Atmosphere-interest degree		16,4	33,6	26,4	17,3	6,2	2	0,26	
Atmosphere-disciplined degree		12,7	39,4	34,0	11,7	2,2	2	0,25	
Atmosphere-motivation level		8,5	29,9	38,5	18,5	4,5	3	0,24	

Table A.7: Descriptive analysis of evaluation data of Lectures course 2003.

entropy.xpl

	0	1	2	3	4	5	Mode	Entropy	Merge
Global overall ratings of the course	9,1	16,7	49,6	18,6	5,5	0,5	2	0,18	4/5
Course missing times		50,3	27,7	13,9	4,9	3,2	1	0,22	
Course missing reason	54,3	3,1	4,8	1,4	6,4	29,9	0	0,22	
explain ability		25,3	50,7	18,2	5,3	0,6	2	0,25	4/5
Content clarity		20,1	50,7	23,6	4,7	1,0	2	0,23	4/5
Transparancy quality		16,2	44,2	30,0	8,1	1,6	2	0,24	4/5
Didactical ability		14,9	43,9	31,2	8,1	1,9	2	0,24	4/5
Stimulation of independent thought		12,8	37,3	34,2	13,2	2,5	2	0,25	
Willingess to answer questions		36,5	47,9	13,1	2,0	0,4	2	0,22	4/5
Quality of answered questions		22,0	52,1	21,5	3,7	0,7	2	0,22	4/5
Time allowed after course	27,5	11,0	30,6	27,1	3,3	0,6	2	0,26	4/5
Aspects covered deepness		17,5	52,7	25,6	3,6	0,6	2	0,23	4/5
Topic structure clarity		20,1	49,5	24,7	4,8	0,8	2	0,23	4/5
Practical example application		18,5	39,0	29,5	11,2	1,7	2	0,25	4/5
Choice of lecture notes		18,7	42,6	27,8	8,9	2,0	2	0,25	4/5
Availability of lecture notes		23,2	44,6	23,7	6,3	2,2	2	0,25	4/5
Presence in the internet		25,2	44,4	21,7	6,5	2,2	2	0,25	4/5
Lecture speed		4,6	26,3	60,4	7,3	1,5	3	0,21	4/5
Mathematical level		6,8	25,6	63,4	3,6	0,6	3	0,19	4/5
Difficulty		6,0	31,6	57,3	4,5	0,7	3	0,19	4/5
Attention span		29,5	45,2	17,9	6,3	1,0	2	0,22	4/5
Knowledge increase		14,8	44,6	29,5	9,6	1,5	2	0,23	4/5
Preparation level		13,1	40,9	26,0	11,6	8,4	2	0,26	
Challenging feeling	8,2	4,3	25,4	57,2	4,4	0,5	3	0,23	4/5
Atmosphere-stress level		22,4	35,3	26,3	13,5	2,5	2	0,26	
Atmosphere-interest degree		12,6	36,7	33,5	13,8	3,4	2	0,26	
Atmosphere-disciplined degree		14,0	42,5	33,8	8,4	1,2	2	0,25	4/5
Atmosphere-motivation level		7,6	31,9	42,7	14,9	2,9	3	0,25	

Table A.8: Descriptive analysis of evaluation data of Exercise course 2003.

entropy.xpl

	0	1	2	3	4	5	Mode	Entropy	Merge
Global overall ratings of the course	10,7	8,2	40,5	26,0	10,5	4,2	2	0,19	
Course missing times		36,9	28,2	18,3	9,9	6,7	1	0,23	
Course missing reason	40,8	6,3	5,2	6,3	7,0	34,4	0	0,24	
Explain ability		15,5	43,3	26,3	9,9	5,0	2	0,27	
Content clarity		11,4	41,9	28,8	12,5	5,4	2	0,25	
Transparancy quality		15,1	36,8	29,3	14,7	4,1	2	0,25	
Didactical ability		12,0	37,8	32,4	11,8	6,1	2	0,24	
Stimulation of independent thought		10,4	34,6	36,1	14,0	4,7	3	0,24	
Willingess to answer questions		23,8	46,3	22,1	6,0	1,8	2	0,23	
Quality of answered questions		14,6	46,3	28,7	7,3	3,1	2	0,22	
Time allowed after course	22,1	9,3	27,5	34,2	5,1	1,9	3	0,25	
Aspects covered deepness		11,0	46,3	30,8	9,3	2,7	2	0,23	
Topic structure clarity		14,7	42,3	27,4	10,6	5,1	2	0,24	
Related topics reference		9,0	39,4	36,8	11,6	3,1	2	0,23	
Practical example application		15,1	38,8	29,3	12,6	4,3	2	0,24	
Choice of lecture notes		15,7	34,9	28,4	15,1	6,0	2	0,25	
Availability of lecture notes		19,3	38,1	24,9	13,1	4,6	2	0,25	
Presence in the internet		21,0	37,9	23,7	12,2	5,1	2	0,25	
Content update		14,7	41,5	35,3	6,7	1,8	2	0,23	
Relevance beween lecture and exercise	32,7	14,0	27,0	17,6	5,7	3,0	0	0,26	
Lecture speed		8,3	30,0	53,5	6,8	1,5	3	0,22	4/5
Mathematical level		13,9	27,2	52,9	4,7	1,2	3	0,21	4/5
Difficulty		10,7	31,3	52,9	4,7	0,4	3	0,20	4/5
Interest degree		18,2	37,5	27,2	12,0	5,2	2	0,23	
Attention span		17,3	41,5	26,1	10,7	4,3	2	0,23	
Knowledge increase		9,1	37,6	33,5	13,9	5,9	2	0,24	
Preparation level		17,5	46,4	20,1	7,9	8,1	2	0,24	
Challenging feeling	13,3	8,1	23,5	48,8	5,4	0,9	3	0,24	4/5
Atmosphere-stress level		20,2	34,7	26,5	13,1	5,5	2	0,25	
Atmosphere-interes degree		10,5	33,1	28,5	17,2	10,6	2	0,26	
Atmosphere-disciplined degree		11,2	37,2	36,5	11,4	3,7	2	0,24	
Atmosphere-motivation level		5,5	25,6	40,9	18,6	9,3	3	0,24	

Table A.9: Descriptive analysis of evaluation data of Lecture course 2002.

entropy.xpl

	0	1	2	3	4	5	Mode	Entropy	Merge
Global overall ratings of the course	11,6	11,1	50,4	20,4	5,2	1,2	2	0,18	4/5
Course missing times		51,8	26,3	13,3	4,6	4,0	1	0,21	
Course missing reason	55,0	3,3	4,7	1,8	4,5	30,7	0	0,21	
Explain ability		20,9	51,2	20,5	6,2	1,3	2	0,25	4/5
Content clarity		15,2	49,9	26,2	7,4	1,3	2	0,23	4/5
Transparancy quality		11,0	42,5	33,8	9,9	2,7	2	0,23	
Didactical ability		11,6	43,4	33,4	8,9	2,8	2	0,23	
Stimulation of independent thought		11,3	37,6	35,5	12,9	2,8	2	0,24	
Willingess to answer questions		32,7	49,7	14,0	3,0	0,5	2	0,21	4/5
Quality of answered questions		16,3	48,4	28,7	5,5	1,2	2	0,22	4/5
Time allowed after course	27,4	10,0	29,9	29,3	2,5	0,9	2	0,25	4/5
Aspects covered deepness		12,2	54,2	27,9	4,6	1,0	2	0,22	4/5
Topic structure clarity		15,6	50,6	24,7	7,9	1,1	2	0,22	4/5
Practical example application		15,0	36,7	31,9	13,3	3,1	2	0,24	
Choice of lecture notes		13,5	38,5	31,2	12,9	3,8	2	0,25	
Availability of lecture notes		19,1	41,7	25,3	10,2	3,8	2	0,25	
Presence in the internet		21,2	44,8	23,2	7,7	3,0	2	0,24	
Lecture speed		6,5	29,1	57,1	6,3	1,0	3	0,21	4/5
Mathematical level		11,8	28,4	56,5	2,9	0,4	3	0,20	4/5
Difficulty		9,1	33,8	52,9	3,9	0,3	3	0,19	4/5
Attention span		28,5	45,3	19,1	5,9	1,1	2	0,21	4/5
Knowledge increase		11,4	45,0	32,2	9,1	2,3	2	0,22	
Preparation level		14,6	42,6	22,9	11,6	8,3	2	0,25	
Challenging feeling	7,8	6,8	27,5	53,2	4,2	0,5	3	0,23	4/5
Atmosphere-stress level		20,2	31,5	27,6	16,3	4,4	2	0,25	
Atmosphere-interest degree		11,4	36,9	33,8	13,7	4,1	2	0,25	
Atmosphere-disciplined degree		11,1	40,5	36,1	10,0	2,3	2	0,24	
Atmosphere-motivation level		7,0	29,2	43,7	15,8	4,2	3	0,24	

Table A.10: Descriptive analysis of evaluation data of Exercise course 2002.

entropy.xpl

Item code	Item meaning
Lecturer	
b1	Explain ability
b2	Content clarity
b3	Transparancy quality
b4	Didactical ability
b5	Stimulation of independent thought
b6	Willingess to answer questions
b6_2	Quality of answered questions
b8	Time allowed after course
Lecture Concept	
c1	Aspects covered deepness
c2	Topic structure clarity
c3	Related topics reference
c4	Practical example application
c5	Choice of lecture notes
c6	Availability of lecture notes
c7	Presence in the internet
c8	Content update
c9	Relevance beween lecture and exercise
Course attributes	
d1	Lecture speed
d2	Mathematical level
d3	Difficulty
Self assesment	
e1	Interest degree
e2	Attention span
e3	Knowledge increase
e4	Preparation level
e5	Challenging feeling
Course atmosphere	
f1	Atmosphere-stress level
f2	Atmosphere-interest degree
f2	Atmosphere-disciplined degree
f4	Atmosphere- motivation level

Table A.11: Code of the response items.

	Lecture 2003	Variance explained	Lecture 2002	Variance explained
λ_1	10,83	0,37	10,70	0,37
λ_2	2,79	0,47	2,90	0,47
λ_3	1,85	0,53	1,92	0,54
λ_4	1,32	0,58	1,43	0,58
λ_5	1,14	0,62	1,11	0,62
λ_6	1,09	0,66	0,94	0,65
λ_7	0,89	0,69	0,81	0,68
λ_8	0,82	0,71	0,80	0,71
λ_9	0,78	0,74	0,76	0,74
λ_{10}	0,73	0,77	0,70	0,76
λ_{11}	0,65	0,79	0,68	0,78
λ_{12}	0,62	0,81	0,65	0,81
λ_{13}	0,52	0,83	0,54	0,82
λ_{14}	0,49	0,85	0,53	0,84
λ_{15}	0,49	0,86	0,48	0,86
λ_{16}	0,44	0,88	0,46	0,88
λ_{17}	0,41	0,89	0,42	0,89
λ_{18}	0,40	0,90	0,39	0,90
λ_{19}	0,37	0,92	0,38	0,92
λ_{20}	0,34	0,93	0,37	0,93
λ_{21}	0,32	0,94	0,33	0,94
λ_{22}	0,30	0,95	0,29	0,95
λ_{23}	0,27	0,96	0,26	0,96
λ_{24}	0,26	0,97	0,24	0,97
λ_{25}	0,23	0,98	0,24	0,98
λ_{26}	0,20	0,98	0,20	0,98
λ_{27}	0,16	0,99	0,19	0,99
λ_{28}	0,16	0,99	0,17	1,00
λ_{29}	0,15	1,00	0,14	1,00
$\lambda > 1$	6 factors		5 factors	

Table A.12: Eigenvalues for Lecture sample correlation matrix.

	Exercise 2003	Variance explained	Exercise 2002	Variance explained
λ_1	8,43	0,34	9,08	0,36
λ_2	3,19	0,46	3,15	0,49
λ_3	1,76	0,54	1,70	0,56
λ_4	1,40	0,59	1,40	0,61
λ_5	1,26	0,64	1,06	0,66
λ_6	0,92	0,68	0,91	0,69
λ_7	0,88	0,71	0,87	0,73
λ_8	0,88	0,75	0,74	0,76
λ_9	0,78	0,78	0,71	0,78
λ_{10}	0,67	0,81	0,68	0,81
λ_{11}	0,54	0,83	0,51	0,83
λ_{12}	0,49	0,85	0,49	0,85
λ_{13}	0,49	0,87	0,46	0,87
λ_{14}	0,45	0,89	0,43	0,89
λ_{15}	0,42	0,90	0,41	0,90
λ_{16}	0,39	0,92	0,38	0,92
λ_{17}	0,35	0,93	0,33	0,93
λ_{18}	0,30	0,94	0,30	0,94
λ_{19}	0,28	0,96	0,26	0,95
λ_{20}	0,24	0,96	0,24	0,96
λ_{21}	0,23	0,97	0,22	0,97
λ_{22}	0,21	0,98	0,21	0,98
λ_{23}	0,20	0,99	0,18	0,99
λ_{24}	0,17	1,00	0,16	0,99
λ_{25}	0,15	1,00	0,15	1,00
$\lambda > 1$	5 factors		5 factors	

Table A.13: Eigenvalues for Exercise sample correlation matrix.

Item	Lecture 2003	Lecture 2002	Exercise 2003	Exercise 2002
Explain ability	1	1*	1*	1*
Content clarity	1	1*	1	1
Transparancy quality	1	1	1	1
Didactical ability	1	1	1*	1*
Stimulation of independent thought				
Willingess to answer questions	4*	4*	1*	1*
Quality of answered questions	4*	4*	1*	1*
Time allowed after course				
Aspects covered deepness	1*	1	4	5
Topic structure clarity	1*	1	4	5
Related topics reference	1		-	-
Practical example application			3	3*
Choice of lecture notes	2	2*	3*	3*
Availability of lecture notes	2*	2*	3*	3*
Presence in the internet	2	2*		
Content update			-	-
Relevance beween lecture and exercise			-	-
Lecture speed	3*	3*	2*	2*
Mathematical level	3	3*	2*	2*
Difficulty	3*	3*	2*	2*
Interest degree	5*	5	-	-
Attention span	5*	5*		4
Knowledge increase	5	5*	5	4
Preparation level				
Challenging feeling	3	3	2*	2*
Atmosphere-stress level				
Atmosphere-interest degree	5	5*	5*	4*
Atmosphere-disciplined degree			5	4
Atmosphere- motivation level	5	5	5*	4*

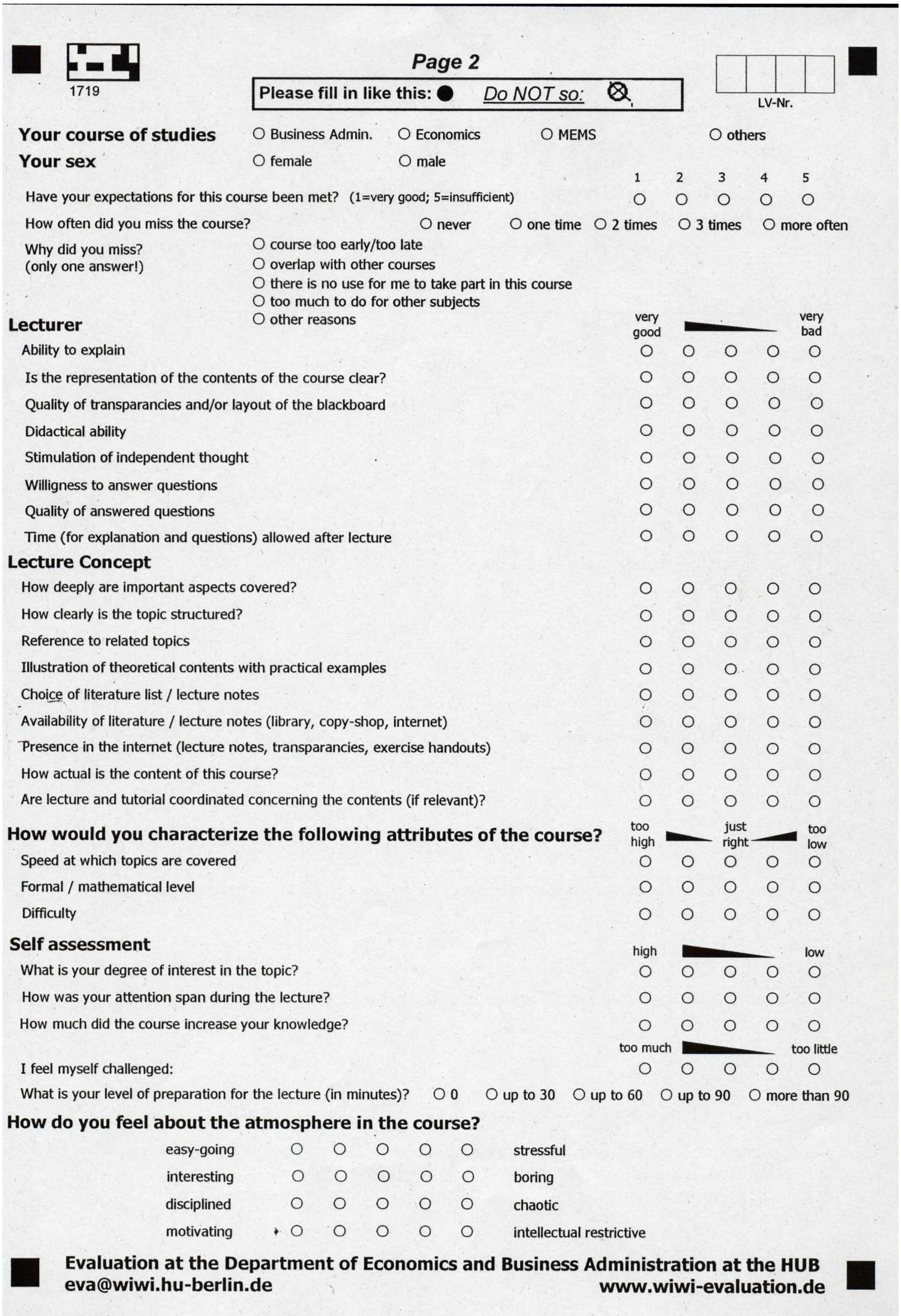
Table A.14: Factor structure of five-factor model.

	SPSS	M-plus	Correlation	Significance level
Lecture course 2003	1	1	0,930	0,000
	2	5	0,980	0,000
	3	2	0,986	0,000
	4	3	-0,994	0,000
	5	6	0,746	0,000
	6	4	0,589	0,001
Lecture course 2002	1	1	0,947	0,000
	2	3	-0,995	0,000
	3	2	0,992	0,000
	4	5	0,938	0,000
	5	4	0,789	0,000
Exercise course 2003	1	1	0,984	0,000
	2	5	0,776	0,000
	3	4	-0,983	0,000
	4	3	0,978	0,000
	5	5	0,614	0,001
Exercise course 2002	1	1	0,966	0,000
	2	2	-0,984	0,000
	3	3	0,988	0,000
	4	4	0,803	0,000
	5	4	0,713	0,000

Table A.15: Correlation coefficients between 5 factor loadings calculated from SPSS and M-plus.

	SPSS	M-plus	Correlation	Significance level	Factor intepretation
Lecture course 2003	1	1	0,996	0,000	Communication Skill & Student reactions
	2	3	0,977	0,000	Lecture notes
	3	2	-0,990	0,000	Course attributes
Lecture course 2002	1	1	0,989	0,000	Communication Skill & Student reactions
	2	2	0,987	0,000	Lecture notes
	3	3	0,996	0,000	Course attributes
Exercise course 2003	1	1	0,996	0,000	Communication Skill & Student reactions
	2	3	0,970	0,000	Course attributes
	3	2	-0,992	0,000	lecture notes
Exercise course 2002	1	1	0,998	0,000	Communication Skill & Student reactions
	2	2	-0,995	0,000	Course attributes
	3	3	0,981	0,000	lecture notes

Table A.16: Correlation coefficients between 3 factor loadings calculated from SPSS and M-plus.



1719

Page 2

Please fill in like this: ● *Do NOT so:* ⊗

LV-Nr.

Your course of studies ○ Business Admin. ○ Economics ○ MEMS ○ others

Your sex ○ female ○ male

	1	2	3	4	5
Have your expectations for this course been met? (1=very good; 5=insufficient)	○	○	○	○	○

How often did you miss the course? ○ never ○ one time ○ 2 times ○ 3 times ○ more often

Why did you miss? (only one answer!)
- ○ course too early/too late
- ○ overlap with other courses
- ○ there is no use for me to take part in this course
- ○ too much to do for other subjects
- ○ other reasons

Lecturer	very good				very bad
Ability to explain	○	○	○	○	○
Is the representation of the contents of the course clear?	○	○	○	○	○
Quality of transparancies and/or layout of the blackboard	○	○	○	○	○
Didactical ability	○	○	○	○	○
Stimulation of independent thought	○	○	○	○	○
Willigness to answer questions	○	○	○	○	○
Quality of answered questions	○	○	○	○	○
Time (for explanation and questions) allowed after lecture	○	○	○	○	○
Lecture Concept					
How deeply are important aspects covered?	○	○	○	○	○
How clearly is the topic structured?	○	○	○	○	○
Reference to related topics	○	○	○	○	○
Illustration of theoretical contents with practical examples	○	○	○	○	○
Choice of literature list / lecture notes	○	○	○	○	○
Availability of literature / lecture notes (library, copy-shop, internet)	○	○	○	○	○
Presence in the internet (lecture notes, transparancies, exercise handouts)	○	○	○	○	○
How actual is the content of this course?	○	○	○	○	○
Are lecture and tutorial coordinated concerning the contents (if relevant)?	○	○	○	○	○

How would you characterize the following attributes of the course?	too high		just right		too low
Speed at which topics are covered	○	○	○	○	○
Formal / mathematical level	○	○	○	○	○
Difficulty	○	○	○	○	○

Self assessment	high				low
What is your degree of interest in the topic?	○	○	○	○	○
How was your attention span during the lecture?	○	○	○	○	○
How much did the course increase your knowledge?	○	○	○	○	○
	too much				too little
I feel myself challenged:	○	○	○	○	○

What is your level of preparation for the lecture (in minutes)? ○ 0 ○ up to 30 ○ up to 60 ○ up to 90 ○ more than 90

How do you feel about the atmosphere in the course?

easy-going	○	○	○	○	○	stressful
interesting	○	○	○	○	○	boring
disciplined	○	○	○	○	○	chaotic
motivating	○	○	○	○	○	intellectual restrictive

Evaluation at the Department of Economics and Business Administration at the HUB
eva@wiwi.hu-berlin.de **www.wiwi-evaluation.de**

Figure A.1: Evaluation form.

Bibliography

Evaluation, P. (2002), Evaluation, sommersemester2002, Technical report, Wirtschaftswissenschaltliche Fakultät der Humboldt-Universität zu Berlin.

Evaluation, P. (2003), Evaluation, sommersemester2003, Technical report, Wirtschaftswissenschaltliche Fakultät der Humboldt-Universität zu Berlin.

Härdle, W., Klinke, S. & Müller, M. (2001), *XploRe: Learning Guide*, Springer.

Härdle, W. & Simar, L. (2003), *Applied Multivariate Statistical Analysis*, Springer.

J.Bartholomew, D., Steele, F., Moustaki, I. & I.Galbraith, J. (2002), *The Analysis and Interpretation of Multivariate Data for Social Scientists*, Chapman & Hall/CRC.

Muthén, B. (1998), *Mplus User's Guide*.

Rönz, B. (1997), *Computergestuetzte Statistik I*, chapter 2.

Rönz, B. (2001), *Generalisierte Lineare Modelle*, chapter 4.

Schafer, J. & Olsen, M. (1997), Multiple imputation for multivariate missing-data problems: a data analyst's perspective, Technical report, Center for Prevention Methodology at Penn State.

Studienordnung für den Diplomstudiengang Betriebswirtschaftslehre 2000, Humboldt Universität zu Berlin (2000).

Studienordnung für den Diplomstudiengang Volkswirtschaftslehre 2000, Humboldt Universität zu Berlin (2000).