HIGHER DIMENSIONAL LOG-LINEAR MODEL AND ITS APPLICATION

Mustofa Usman[1*], Widiarti[1], Rudi Ruswandi[1], Faiz AM Elfaki[2] and Jamal I Daoud[2]
[1]Department of Mathematics, Faculty of Sciences and Mathematics, University of Lampung, Indonesia
[2]Department of Sciences, Faculty of Engineering, IIUM, Malaysia
*Email: usman_alfha@yahoo.com

ABSTRACT*: In this study the model of higher dimensional log-linear model is applied to four categorical variables in Education. The data are collected from the alumni data of University of Lampung, from 2010 to 2013 and about 9060 alumni involved. In this study, the variables of interest are: Length of Study with three categories (<4.5 years; 4.5- 5.5 years; and >5.5 years), Field of Study with three categories (Sciences, Social Sciences, and Education), Sex with two categories (Male, and Female), GPA in scale 0 to 4 with three categories (<3.0, 3.0-3.5, and >3.5). In this study the aims are going to find the best model to explain the relationship among the factors. By using hierarchical Log-linear Model Analysis and backward method it was found that the best model for the data with three variables interactions in the model are: Length of Study*SEX*GPA, Length of Study*Sciences*GPA, and Sex*Sciences*GPA.*

Key words*: log-linear models; categorical data; interactions; backward method.*

INTRODUCTION

A new data analysis technique known as log-linear models has been developed over the past decade, providing a means for the analysis of qualitative data at a level of sophistication that has long been available for quantitative data. Under the log-linear procedures, a researcher can establish a linear model for the observed frequencies in the cells of a multidimensional contingency table in a manner similar to that used in the analysis of variance [1, 2, 3, 4]. The log-linear models methodology arose primarily within the context of survey research where the interest was in understanding the interrelationships among qualitative variables used to define a multidimensional contingency table [1]. The development and application of methods for analyzing categorical data in many fields of study such as in medical sciences, epidemiology, economics, social science, education and others are very extensive. There are wide literature and research papers on log-linear model in the last forty years [3, 5, 6, 7, 8]. The application of log-linear model in education we can found such as [9] whom discussed how to increase satisfaction with online learning. Ting and Abella [10] used log-linear model to measuring student course evaluations. The application of log-linear model in evaluation of education and Rasch Model Test can be found in [11, 12]. Analysis Test results in education [1], one form of analysis employed in test norming is to compare the test performance of subgroups of interest. Historically, this has been done via tests of equality of group means and/or variances as well as goodness-of-fit tests between pairs of group distributions. The log-linear model approach, however, enables the simultaneous testing of the homogeneity of entire test score distributions for multiple groups. Fienberg [8] based on educational data given by Beaton [13] give an example how to analyze three dimensional categorical data by using log-linear models.

The aims of this study are going to analyze the interrelation among four categorical education data, namely, Length of Study with three categories (<4.5 years; 4.5- 5.5 years; and >5.5 years), Field of Study with three categories (Sciences, Social Sciences, and Education), Sex with two categories (Male, and Female), GPA in scale 0 to 4 with three categories (<3.0, 3.0-3.5, and > 3.5). In this study the log-linear model will be applied to analysis four dimensional categorical data. And the best model will be used to explain the relationship among the four dimensional categorical data.

GENERAL LOG-LINEAR MODEL AND TESTING

Haberman [14] presented general log-linear model that specifies the relations among a set of observable categorical variables. The models explain the structure of the contingency table that is formed by cross-classifying the set of variables of interest. This is accomplished by specifying a linear decomposition of the natural log of expected contingency table frequencies. In higher dimensional table, some complications arise due to the number of possible association and interaction terms, making model selection more difficult. For four dimensional tables for this study, the factors are given in the following table.

Table 1. Factors and Categories

Factors	Categories		
Length of Study (L)	<4.5 years	4.5—5.5 years	>5.5 years
Field of Study(F)	Science	Social Science	Education
Sex(S)	Male	Female	
GPA(G)	< 3.0	3.0 – 3.5	>3.5

Following Agresti [7] and Christensen [15], some possible models are:

Saturated model

$$\log(m_{ijkl}) = \lambda + \lambda_i^L + \lambda_j^F + \lambda_k^S + \lambda_l^G + \lambda_{ij}^{LF} + \lambda_{ik}^{LS} + \lambda_{il}^{LG} + \lambda_{jk}^{FS} + \lambda_{jl}^{FG} + \lambda_{kl}^{SG} + \lambda_{ijk}^{LFS} + \lambda_{ijl}^{LFG} + \lambda_{ikl}^{LSG} + \lambda_{jkl}^{FSG} + \lambda_{ijkl}^{LFSG} \quad (1)$$

3-way interaction model

$$\log(m_{ijkl}) = \lambda + \lambda_i^L + \lambda_j^F + \lambda_k^S + \lambda_l^G + \lambda_{ij}^{LF} + \lambda_{ik}^{LS} + \lambda_{il}^{LG} + \lambda_{jk}^{FS} + \lambda_{jl}^{FG} + \lambda_{kl}^{SG} + \lambda_{ijk}^{LFS} + \lambda_{ijl}^{LFG} + \lambda_{ikl}^{LSG} + \lambda_{jkl}^{FSG} \quad (2)$$

2-way interaction model

$$\log(m_{ijkl}) = \lambda + \lambda_i^L + \lambda_j^F + \lambda_k^S + \lambda_l^G + \lambda_{ij}^{LF} + \lambda_{ik}^{LS} + \lambda_{il}^{LG} + \lambda_{jk}^{FS} + \lambda_{jl}^{FG} + \lambda_{kl}^{SG} \quad (3)$$

Independent model

$$\log(m_{ijkl}) = \lambda + \lambda_i^L + \lambda_j^F + \lambda_k^S + \lambda_l^G \quad (4)$$

Saturated model always provides a perfect fit of the data. However, smaller models have more powerful interpretations and are often better predictive tools than large models.

It is common that model (1),(2),(3) and (4) can be written respectively as

$$[\,L\,F\,S\,G\,],$$
$$[LFS]\,[LFG]\,[LSG\,][FSG],$$
$$[LF][LS][LG][FS][FG][SG], \qquad (5)$$

and

$$[L][F][S][G].$$

To test each of the model and to test each model against the saturated models. First we determine the expected count for each model (1),(2),(3), and (4) by using maximum likelihood estimation [8, 15, 16]. The expected count for model (1), (2), (3) and (4) are respectively:

$$\hat{m}^{(1)}_{ijkl} = n_{ijkl}\,, \qquad (6)$$

$$\hat{m}^{(2)}_{ijkl} = n_{\bullet\bullet\bullet\bullet}\,\hat{p}_{ijkl} = \frac{n_{ijk\bullet}\, n_{ij\bullet l}\, n_{i\bullet kl}\, n_{\bullet jkl}\,(n_{\bullet\bullet\bullet\bullet})^2}{n_{ij\bullet\bullet} n_{i\bullet k\bullet} n_{i\bullet\bullet l} n_{\bullet jk\bullet} n_{\bullet j\bullet l} n_{\bullet\bullet kl}}, \quad (7)$$

$$\hat{m}^{(3)}_{ijkl} = n_{\bullet\bullet\bullet\bullet}\hat{p}_{ijkl} = \frac{n_{ij\bullet\bullet} n_{i\bullet k\bullet} n_{i\bullet\bullet l} n_{\bullet jk\bullet} n_{\bullet j\bullet l}\, n_{\bullet\bullet kl}}{n_{i\bullet\bullet\bullet} n_{\bullet j\bullet\bullet} n_{\bullet\bullet k\bullet} n_{\bullet\bullet\bullet l} n_{\bullet\bullet\bullet\bullet}}, \quad (8)$$

and

$$\hat{m}^{(4)}_{ijkl} = n_{\bullet\bullet\bullet\bullet}\hat{p}_{ijkl} = \frac{n_{i\bullet\bullet\bullet}\, n_{\bullet j\bullet\bullet}\, n_{\bullet\bullet k\bullet}\, n_{\bullet\bullet\bullet l}}{(n_{\bullet\bullet\bullet\bullet})^3}\,. \qquad (9)$$

Where i=1,2,..., I; j=1,2..,J ; k=1,2,...,K; and l=1,2,...,L.

The symbol $\bullet$ means the summation over the corresponding index [16], for example $n_{ij\bullet\bullet} = \sum_k \sum_l n_{ijkl}$.

To test the respective models, model (1), (2), (3), and (4) we can use the Pearson chi-square test statistic

$$\chi^2 = \sum_i^I \sum_j^J \sum_k^K \sum_l^L \frac{(n_{ijkl} - \hat{m}^{(s)}_{ijkl})^2}{\hat{m}^{(s)}_{ijkl}}\,, \qquad (10)$$

where s=1,2,3,4, or by likelihood ratio test statistic

$$G^2 = 2\sum_i^I \sum_j^J \sum_k^K \sum_l^L n_{ijkl}\log(n_{ijkl}/\hat{m}^{(s)}_{ijkl})\,. \qquad (11)$$

The degrees of freedom [8] for each model are given below:

Df for model (1), saturated model,

$$df = 0,$$

Df for model (2), 3-way interaction model

$$df = [(I-1)(J-1)(K-1)(L-1)],$$

Df for model (3), 2-way interaction model

$$df = [IJK-IJ-IK-IL-JK-JL-KL+2(I+J+K+L)-3],$$

Df for model (4), independent model

$$df = [\, IJKL- I-J-K-L +3\,].$$

To test for comparison between the models [15] for example the likelihood ratio test statistics for testing model (2) vs. model (1), saturated model, the test is

$$G^2(2 \text{ vs. } 1) = 2\sum_{ijkl} n_{ijkl}\log(n_{ijkl}/\hat{m}^{(2)}_{ijkl}) \qquad (12)$$

and to test between model (r) and (s), the likelihood ratio test is

$$G^2(r \text{ vs. } s) = 2\sum_{ijkl}\hat{m}^{(s)}_{ijkl}\log(\hat{m}^{(s)}_{ijkl}/\hat{m}^{(r)}_{ijkl}) \quad (13)$$

In a simple form

$$G^2\,(r \text{ vs. } s) = G^2(r \text{ vs. } 1) - G^2(s \text{ vs. } 1) \qquad (14)$$

With the degrees of freedom for the test is

$$df(r \text{ vs. } s) = df\,(r \text{ vs. } 1) - df(s \text{ vs. } 1). \qquad (15)$$

The methods of obtaining G^2(r vs. s) and df (r vs. s) from G^2's and df's for testing against saturated models are basic to log-linear model practice [15, 16]. To find the best model in this study we will use Akaike's Information Criterion and Backward Method. Akaike [17] proposed a criterion of the information contained in a statistical models. He advocated choosing the model that maximizes this information. For log-linear model, AIC criterion to choosing a model, say X, that minimizes

$$A_X = G^2(X) - |q-2r|, \qquad (16)$$

where r is the df for X model, and q is df for saturated model, i.e. q is cell in the table. The Backward elimination procedure is based on comparing models and does not consider whether the reduce models fit relative to the saturated model. In backward procedure, we will start with the most complex model, which in this case would be all three factor model {LSF, LFG, LSG, FSG}. We will used cut off point of α=0.05 as our criteria of deleting. At each stage of our selection, we delete the term for which the p-value will be least significant (p-value > 0.05).

DATA AND LOG-LINEAR MODELS ANALYSIS

The following data are from undergraduate alumni University of Lampung from 2010-2013.

Table 2. Data Length of Study(L), Field of Study (F), Sex (S) and GPA (G) undergraduate alumni University of Lampung 2010- 2013.

Length of Study	Field of Study	Sex	GPA <3.0	GPA 3.0-3.5	GPA >3.5
<4.5 years	Science	M	17	600	136
		F	13	755	451
	Social Science	M	43	189	44
		F	86	461	119
	Education	M	10	193	67
		F	42	909	258
4.5-5.5 years	Science	M	41	301	30
		F	40	196	26
	Social Science	M	165	364	28
		F	247	423	43
	Education	M	36	165	13
		F	105	407	36
>5.5 years	Science	M	136	205	8
		F	33	92	8
	Social Science	M	535	246	4
		F	233	128	2
	Education	M	72	100	3
		F	89	104	9

Source: University of Lampung, Data Alumni 2010-2013.

There are about 48.48% students that can finish their study on time , namely < 4.5 years, with the modes of GPA in the range 3.0-3.5 are about 34.29% (female students 23.45 %, male students 10.84 %) and about 11.86% students got GPA above 3.5 (female students 9.13% and male students 2.73%), and only 2.32% students finished on time with GPA <3.0. There are about 29.42% students that can finish their study for 4.5-5.5 years, with the modes of GPA in the range 3.0-

3.5 about 20.48% (female students 11.32%, male students 9.16%) and about 1.94% students got GPA above 3.5 (female students 1.16% and male students 0.78%), and only 7.22% students finished with GPA <3.0. There are about 22.10% students that can finish their study for more than 5.5 years, with the modes of GPA in the range 3.0-3.5 about 9.65% (female students 3.57%, male students 6.08%) and about 0.37% students got GPA above 3.5 (female students 0.17% and male students 0.16%), and only 12.11% students finished with GPA <3.0.

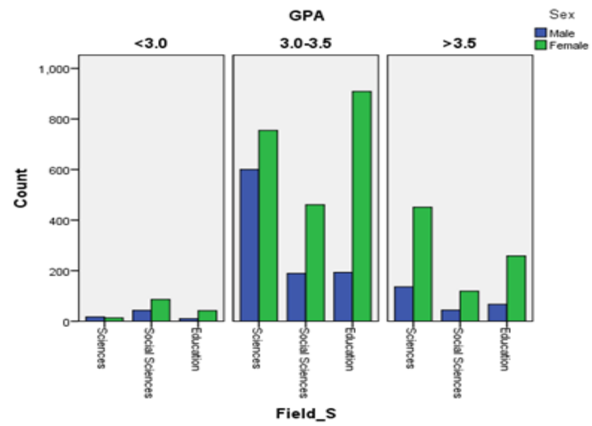


Fig.1. Graphic data GPA, Sex and Field of Study who graduated <4.5 years

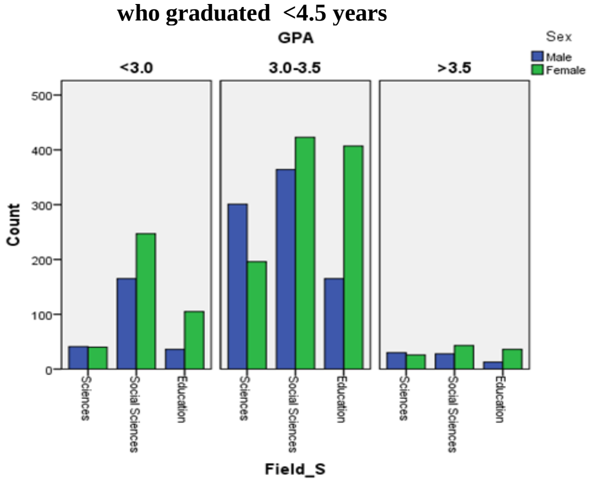


Fig.2. Graphic data GPA, Sex and Field of Study

From the results of analysis of the model (1), (2), (3), and (4), model (2) based on the p-value and the minimum of AIC is the best model among the four models. The parameter estimates and testing the parameters based on model (2) is given below:

From log linear model analysis by using SAS, it was found that for the models given in (1), (2), (3) and (4) the results are as follow:

From Table 4, the interaction L*S*F is not significant (p-value=0.1725), so as suggested by the backward method we can delete the term which is not significant. The new model we found then

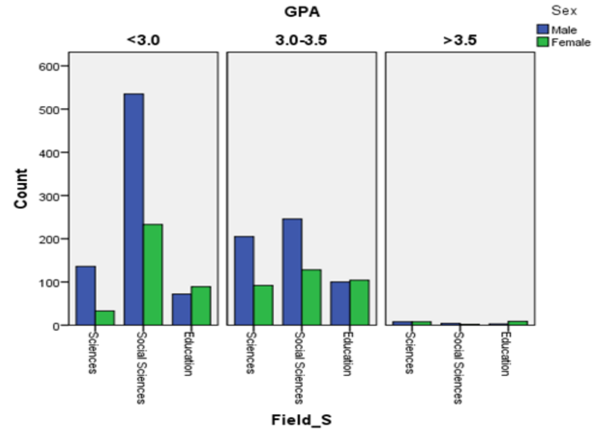


Fig.3. Graphic data GPA, Sex and Field of Study who graduated more than 5.5 years.

Table 3. Log-linear model analysis

Model	df	Likelihood Ratio Test	p-value	AIC
[LFSG]	0	-	-	
[LGS][LGF][GSF][LSF]	8	13.39	0.0902	- 24.31
[LG][LS][LF][GS][GF][SF]	28	110.15	< 0.0001	108.15
[L] [F] [S] [G]	46	5237.99	< 0.0001	5199.99

Table 4. Maximum Likelihood Analysis of Variance for Model (2)

Source	df	Chi-Square	p-value
L	2	178.54	<0.0001
F	2	35.86	<0.0001
S	1	39.92	<0.0001
G	2	2030.46	<0.0001
L*G	4	1045.18	<0.0001
L*S	2	100.73	<0.0001
L*F	4	98.01	<0.0001
G*S	2	9.58	0.0083
G*F	4	390.71	<0.0001
S*F	2	138.84	<0.0001
L*G*S	4	13.68	0.0084
L*G*F	8	30.56	0.0002
G*S*F	4	36.52	<0.0001
L*S*F	4	6.38	0.1725
Likelihood Ratio	8	13.69	0.0902

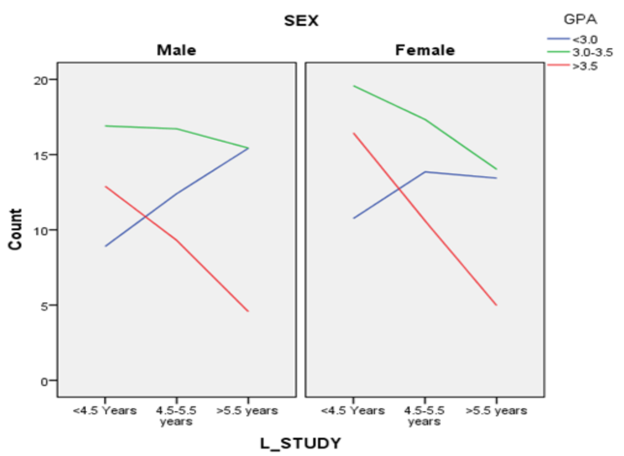
$$\log(m_{ijkl}) = \lambda + \lambda_i^{L} + \lambda_j^{F} + \lambda_k^{S} + \lambda_l^{G} + \lambda_{ij}^{LF} + \lambda_{ik}^{LS} + \lambda_{il}^{LG} + \lambda_{jk}^{FS} + \lambda_{jl}^{FG} + \lambda_{kl}^{SG} + \lambda_{ijl}^{LFG} + \lambda_{ikl}^{LSG} + \lambda_{jkl}^{FSG} \quad (17)$$

In this model all terms are significant and from likelihood ratio test the model(17) fit with the data(p-value>0.05). The Likelihood Ratio test is 20.02 with df=12 and p-value=0.0667 (Table 5). The maksimum likelihood analysis of variance given in Table 4. In this model, there are three ways of interaction among the factors: Length of study, Sex and GPA; Length of study, Field of study and GPA; and Field of study, Sex and GPA. The graph for the interaction Length of study, Sex and GPA; Length of study, Field of study and GPA; and Field of study, Sex and GPA are given in Fig. 4, Fig. 5 and Fig. 6 respectively.

Table 5. Maximum Likelihood Analysis of

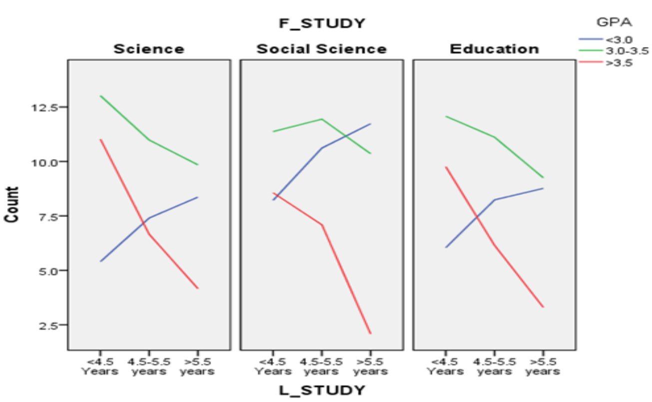
Variance for Model (17)

Source	df	Chi-Square	p-value
L	2	181.47	<0.0001
F	2	38.33	<0.0001
S	1	42.14	<0.0001
G	2	2036.99	<0.0001
L*G	4	1036.00	<0.0001
L*S	2	104.44	<0.0001
L*F	4	97.63	<0.0001
G*S	2	10.17	0.0062
G*F	4	395.14	<0.0001
S*F	2	152.65	<0.0001
L*G*S	4	17.87	0.0013
L*G*F	8	31.07	0.0001
G*S*F	4	36.81	<0.0001
Likelihood Ratio	12	20.02	0.0667



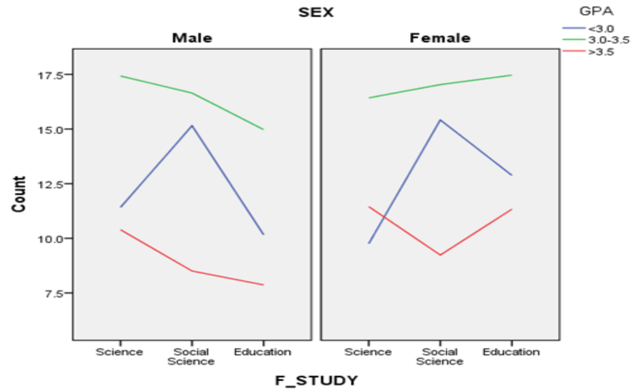
(a) (b)

Fig. 4. Interaction among Length of Study, Sex and GPA



(a) (b) (c)

Fig. 5. Interaction among Length of Study, Field of Study and GPA



(a) (b)

Fig. 6. Interaction among Field of Study, GPA and Sex

The plot for interaction among the factors: Length of study, Sex and GPA is given in Fig. 4. From the graph it was shown that all three curves are not parallel. The curve for GPA <3.0 in Fig. 4(a) and (b) are clearly the main source of interaction. In Fig.4 (a), Male students the GPA 3.0-3.5 is nearly horizontal, this indicates that the number of male students who got GPA 3.0-3.5 across the length of study relatively the same. In Fig.4 (b), Female students the main source of interaction is GPA <3.0 and all three curves are not parallel. The curves for GPA 3.0-3.5 and GPA >3.5 have negative trend across the length of study this means that the longer the students stay in university, the lesser the number of students who got GPA 3.0-3.5 and GPA >3.5.

The plot for interaction among the factors: Length of study, Field of study and GPA is given in Fig. 5.(a), (b) and (c). From the graph it was shown that all three curves are not parallel. In all the field of studies, most of students graduated with GPA 3.0-3.5. In all three plots, the GPA <3.0 has positive trend, and in the Field of social sciences most of students graduated more than 5.5 years, while the GPA 3.0-3.5 and GPA >3.5 have negative trend. Graph also indicated that the GPA <3.0 clearly as the main source of interaction across the length of study and field of studies.

The plot for interaction among the factors: Field of study, Sex and GPA is given in Fig. 6(a) and 6(b). From the graph it was shown that all three curves are really not parallel. In all the field of studies, most of students Male and female, graduated with GPA 3.0-3.5. In groups of students who got the GPA <3.0, the social students has higher frequency compared to others field of study. Graph also indicated that all the factors Field of study, Sex and GPA are as the main source of interaction.

ACKNOWLEDGEMENT:

The authors would like to thank to the University of Lampung for providing the data of the undergraduate alumni for this research.

REFERENCES

[1] Baker, F.B., and Subkoviak,M.J., (1981). Analysis of Test Results via Log-Linear Models, *Aplied Psychological Measurement*, Vol.5, No.4, pp.503-515.

[2] Haberman,S.J.(1973).Loglinear models for frequency data: Sufficient statistics and likelihood equations. *Ann.Math.Statist*, **1**. 617-632.

[3] Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis*. Massachussetts: The MIT Press.

[4] Andersen, A.H. (1974). Multi-dimensional Contingency tables. *Scand. J. Statist*. 1,115-127.

[5] Benedetti,J.K. and Brown.M B.,(1978). Strategies for the selection of log-linear models. *Biometrics,*34, 680-686.

[6] Agresti, A.(1984). *Analysis of Ordinal Categorical Data*, New York: John Wiley.

[7] Agresti, A. (2002). *Categorical data Analysis*. Sec. Ed., New York: John Wiley.

[8] Fienberg, S.E.,(1987). *The Analysis of Cross-Classified Categorical Data*, 2nd Ed. Cambridge, The MIT Press.

[9] Pak, R.J., (2011). How to Increase Satisfaction Learning: Technical Suggestions to enhance the usefulness of importance- performance analysis. *2nd International Conference on Education and Management Technology,* IPEDR vol.13., Singapore.

[10] Ting, D.H., and Abella, M.S.,(2007). Measuring Student Course Evaluations: The use of a loglinear Model. *International Education Journal*, 8(1), 194-204.

[11] Kelderman, H., and Macready,G.B.(1990). The use of Loglinear Models for Assesing Differential Item Functioning Across Manifest and Latent Examinee Groups. *Journal of Education Measurement,* vol.27, No.4, pp.307-327.

[12] Kelderman,H.(1984). Loglinear Rasch Model tests. *Psychometrika,* 49, 223-245.

[13] Beaton,A.E.(1975). The influence of education and ability on salary and attitudes. In F.T. Juster (ed). *Education, Income and Human Behavior*, p.365-396. New York: McGraw-Hill.

[14] Haberman, S.J.(1979). *Analysis of qualitative data: vol.II. New Developments*. New York: Academic Press.

[15] Christensen, R. (1997). *Log-Linear Models and Logistic Regression*. Sec.Ed. New York: Springer-Verlag, Inc.

[16] Christensen, R. (1990). *Log-Linear Model*. New York: Springer-Verlag, Inc.

[17] Akaike, Hirotugu (1973). Information Theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information*, edited by B.N. Petrov and F. Czaki. Budapest: Akademiai Kiado.