Martínez Gómez, M.; Carot Sierra, JM.; Jabaloyes Vivas, JM.; Zarzo Castelló, M. (2011). A multivariate method for analyzing and improving the use of student evaluation of teaching questionnaires: a case study. Quality and Quantity. 45(6):1415-1427. doi:10.1007/s11135-010-9345-5.

Article type: Research article.

Title: A multivariate method for analyzing and improving the use of student evaluation of teaching questionnaires. A case study.

Authors: Mónica Martínez-Gómez[1]; Jose Miguel Carot Sierra; José Jabaloyes; Manuel Zarzo

Affiliation: Department of Applied Statistics, Operations Research and Quality

Universidad Politécnica de Valencia

Camino de Vera s/n, Edificio 7-A

cx46022 VALENCIA (SPAIN)

[1]Corresponding Author: Mónica Martínez-Gómez

**momargo@eio.upv.es**

Universidad Politécnica de Valencia

Camino de Vera s/n, Edificio 7-A

46022 VALENCIA (ESPAÑA)

Tel. +34 963877490

Fax. +34 963877499

## Abstract

Student evaluation of teaching (SET) questionnaires are the most common methods of evaluation used by European universities to assess the quality of teaching delivered by their lecturers. A series of multivariate statistical methods were applied to analyze the underlying structure of the SET questionnaire used by the Universidad Politecnica de Valencia (UPV) in order to develop an appropriate methodology for extracting, analyzing, and interpreting the information contained in the questionnaire. In a first step, a Confirmatory Factorial Analysis (CFA) was developed in order to evaluate the reliability, validity and dimensionality of it, by means of two relatively new parameters commonly used in structural equation modelling: the compound reliability and extracted variance for each latent construct. In a second step, Cluster Analysis (CA) was used to test the ability of the questionnaire for the identification of different categories of lecturers. In the last step, a tree classification method, the Chi-Squared Automatic Interaction Detector (CHAID), was used in order to characterize the different lecturer's categories obtained with CA according to all available information regarding the teaching staff and subjects.

**Keywords:** Quality Indicator, Confirmatory Factor Analysis, Student Questionnaire for Teaching Assessement, Cluster Analysis, CHAID.

## 1. Introduction

The main challenge that must face European universities at the moment is to increase the quality of education, as required by the European Higher Education Area. The process of academic quality improvement necessarily involves the evaluation of teaching staff, as it is an important element to develop a suitable culture of internal evaluation at universities.

There are many ways of evaluating educational activity and therefore, the teaching staff. Berk (2005), in a recent review, describes up to 12 varieties of evaluation. Nevertheless, the form most used in Spain for the evaluation of lecturers is by means of student surveys. But the evaluation of teaching quality is a highly complex process given that this concept is both subjective and multidimensional (Marsh, 1987). In this sense, one of the main weaknesses of the Spanish university system in the area of quality evaluation is the lack of information on data collection tools (Cajide, 1994; González Such, 1997; Rodríguez Sabiote and Gutiérrez Pérez, 2003).

The UPV recognizes the need to design an appropriate statistical methodology for extracting, analyzing, and interpreting the data contained in the SET questionnaire, in order to provide the teaching staff with excellent information to undertake performances of improvement. Similar studies reported in the literature revealed that the underlying dimensions of a particular questionnaire of student evaluations are not always the same and depend on the structure of the questionnaire. Marsh (1982) reported 9 dimensions of teaching effectiveness after analyzing SEEQ (Students Evaluation of Educational Quality) questionnaire.

One major shortcoming of current procedures is that it is uncertain if the survey items of SET questionnaires properly represent the underlying constructs for which they were developed. Marsh (1984) suggested that effective teaching is a hypothetical construct for which there is not a single indicator and the validity of the dimensions of student evaluations should be demonstrated through a construct validation. Gursoy and Umbreit (2005) have argued that the issues of validity regarding student evaluation of teaching effectiveness are highly complex and controversial and also reported contradictory findings about the reliability and the dimensions that represent teaching effectiveness.

The present work is part of research conducted by UPV in the framework of the continuous improvement of teaching quality. The following particular aims have been established for our work: to evaluate reliability, validity, and multidimensionality of the questionnaire; to test its ability to identify clusters of lecturers with a similar teaching quality and to characterize those teacher typologies according to descriptive characteristics related to subjects and lecturers.

## 2. Method

### 2.1 INSTRUMENT

The questionnaire currently used by the UPV since 1994, was created by the Institute of Education Sciences. It was a modification of the original questionnaire.

The first application of the questionnaire was in the year 1987-1988, and it contained 34 items. This number was reduced to 26 in course 1992-1993 and finally to 19 in 1994, because items that were not related to any factorial structure were eliminated and Jackson et al., (1999) have also suggested that an excessive number of items becomes tedious and may influence the intended purpose of the questionnaire. The current version gathers_information of about subject, the evaluated lecturer, the student and, finally, displays a set of 19 items measured on a 5-point Likert scale, anchored with *strongly disagree* at the low end (value of one) and *strongly agree* at the high end (value of 5), with an additional optionof “no opinion”. We refer to these items as R1 to R19 for the statistical analysis. The last item is a criterion or general item which serves to verify the suitability of the questionnaire. A complete list of these items is shown in table 1.

TABLE 1

### 2.2 SAMPLE AND PROCEDURE

We used 3 datasets in the present work. The first one, the database of Individual Scores (IS), is comprised of 19 variables, each corresponding to one individual Likert-scale score for questionnaire items  produced by a single student for a certain subject taught by a lecturer in a specific academic course. The second dataset contains the mean scores for each of the 19 items corresponding to every possible subject-teacher combination, since each subject was not always taught by the same lecturers every year. For this reason, the results from different years cannot be directly compared. These mean scores were obtained by applying a linear transformation to averaged Likert's scores, in order to convert them TO a 0-10 scale. This matrix was labelled as dataset of Mean Scores (MS). Variables were labelled as "mean $R_i$", where i was the item number. Data was comprised of 6 years of questionnaires, from the academic course 1995 to 2001. The total number of questionnaires completed for each period is shown in table 2. In order to facilitate the subsequent characterization of categories, we included all available information about the lecturer (age, category, full-time / part-time employment, doctoral qualifications) for each observation and the subject (department, course from first to sixth, semester). This dataset was labelled as Matrix of Descriptors (MD).

### *2.2.1 Reliability, validity, and multidimensionality of SET*

In order to evaluate whether the measurement scales are suitable at present, it is necessary to study their reliability and validity. Nowadays all studies to assess reliability and validity of questionnaires are based on exploratory factor analysis (EFA) and Cronbach's alpha test for each indicator. However, Batista et al., (2004) have argued that the binomial EFA and Cronbach's alpha test are insufficient to guarantee the reliability and validity of SET questionnaire and they propose Confirmatory Factor Analysis (CFA) as an alternative. Besides, Cronbach's alpha test presents the disadvantage of assuming that each construct presents unidimensionality instead of ensuring it (Hair et al., 1995).

Recent studies have applied multilevel modeling to SET (Marsh and Hattie, 2002; Tings, 2000; Toland and De Ayala, 2005) and other works have been carried out using models estimated through ordinary statistical techniques (Lalla et al., 2004; Göb et al., 2007). However, these methods do not guarantee an optimal solution, and CFA provides a more rigorous test of convergent and discriminant validity than the more traditional multitrait-multimethod analysis (Campbell and Fiske, 1959).

As an alternative, an approach to assess validity and reliability within the framework of confirmatory analysis is used in this work: the indexes compound reliability and the extracted variance. Appropriate values for them are those that exceed .7 and .5, respectively. So a CFA to obtain these reliability parameters was developed, and the underlying structure of the questionnaire was analyzed by testing five structural models, from one to five factors, and considering orthogonal as well as correlated factors to avoid the traditional shortcomings.

In all models, we have used the 19 items of the questionnaire considering that they were correlated with the corresponding latent variables according to each model to avoid specification errors. We followed the criterion of Maximum Likelihood Estimation (ML) under the assumption of multivariate normality, where the loading matrix is the inverse of the implied matrix, $W=(S(p)\otimes \Sigma(p))^{-1}$ (Batista and Coenders, 2000). ML is appropriate when there are missing data, as it is our case. The estimation of the different models was carried out using EQS 6.1.

In structural equation modelling, the goodness-of-fit of a covariance structure model is evaluated by various methods with fit indexes. As the statistical $\chi^2$ is very sensitive to deviations of normality and sample size (Bollen and Long, 1993), other indexes have been considered (Bentler, 1990; Bentler and Bonnet, 1980; Bollen, 1989; Browne and Cudek, 1993; Hu and Bentler, 1999; Jöreskog, 1993; Jöreskog and Sörborm 1981, 1986; Mulaik et al.,1989): the Goodness of Fit Index (GFI) and two variants of it, the Adjusted Goodness of Fit Index (AGFI) and the Parsimony Goodness of Fit Index (PGFI); the Normed Fit Index (NFI); the Not-Normed Adjustment Fit Index (NNFI); the Root Mean Square Error of Approximation (RMSEA); HOELTER and the Index of Crossed Validation (ECVI).

All these analyses were conducted with the IS matrix, selecting the observations corresponding to the academic year 2000-2001 which was the latest and most complete period.

### *2.2.2 Identification of lecturer's categories*

We used BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies), as a cluster analysis improved by Chiu et al. (2001), in order to test the ability of the questionnaire to identify of different categories of lecturers based on the quality of teaching as perceived by students. The advantage of BIRCH is that it works particularly well with large datasets and can handle both categorical and continuous variables simultaneously.

This cluster analysis was conducted with the MS dataset in order to obtain homogenous groups of teacher-subjects with similar characteristics of perceived teaching quality. If BIRCH had been applied to the IS database, we would have obtained clusters of students with similar answer categories, which is not of interest for our study. We selected 16 items for the analysis as the majority of students chose the answer *"I do have enough information to satisfactorily answer this question"* in items R13, R14, and R15 since the questionnaires were given out before the students had been evaluated. We started by analyzing data from the academic year 1999-2000 and then the results were compared with those of 2000-2001, in order to check whether the categories remained stable.

### *2.2.3 Characterization of lecturer categories*

Finally, the different  categories of lecturers obtained with BIRCH were characterized according to all available information included in the MD matrix, regarding the teaching staff and subjects. We used Chi-Squared Automatic Interaction Detector (CHAID) which is a tree classification method originally proposed by Kass (1980) that enables us to segment the dataset according to the most significant predictor variables, as well as to establish groups of lecturers that belong to each category.

We used a new categorical variable called cluster which indicated the cluster to which each observation belongs according to BIRCH. Our aim was to segment this new variable cluster and to establish homogenous groups of subject-lecturers based on the variables of the MD. In all cases, the stopping rule for determining the most convenient splitting degree was fixed at a maximum depth of five splits under the root node. A further recursive splitting would be of little use in our situation because terminal nodes would contain a very small number of cases.

## 3. Results

The application of EFA leads to the identification of five relevant factors that explain 77.95 percent of the overall data variance. Table 3 shows the structure matrix, with the variable loadings (correlation coefficients, *r*) for each of the five factors. As this table shows, the highest loadings (r > .75) for component 1 correspond to the first six items. Representative items are R1 and R2. This dimension could be labelled “Command, organization, and clarity of subject and programme”. Although item R19 was expected to be related to all components, it presents a stronger correlation with component 1, so this factor could be considered the most relevant of the questionnaire to evaluate teaching quality, as is further discussed below.

TABLE 3

Following the same criterion of highest loadings for each component, we could label: Component 2 as “Evaluation”, Component 3 as “Lecturer-student interrelation”, Component 4 as “Resources” and finally, component 5 as “Interrelation with other subjects”.

It appears that the initial structure of the questionnaire is kept, with five factors and an additional item to indicate the overall satisfaction.

The results to verify the reliability and validity of the scales are shown in tables 4 and 5. All values in table 4 are higher than 0.9 which suggests that the SET questionnaire scales have

TABLE 4
TABLE 5

suitable values of internal consistency. On the other hand, only the model hypothesized with five factors display accepted values with respect to the reliability of construct and the variance extracted (table 5).

Table 6 shows the main goodness-of-fit indicators for the different models proposed. Analyzing these indices, we can conclude that the hypothesis to represent the factorial structure of the SET questionnaire by means of a single dimension seems quite inadequate. So, it is necessary to verify the number of dimensions that better represents it. Model 4, with five correlated factors, presents a better fit than the model with two factors, although it continues denoting divergences between the matrix of variances - covariances of the sample and the matrix generated from the model ($\chi^2$= 80254.4, *df* = 139 and p < .05). These divergences are lower in model 5 ($\chi^2$= 43683.4, *df* = 135 and p < .05), and the goodness of fit indexes improve enough, being close to the values recommended for each index, particularly in the case of RMSEA, PNFI and PGFI. This fact indicates that model 5 has the best parsimony, and this is the concept that we must consider when comparing alternative models (Mulaik et al., 1989).

TABLE 6

Although in models, 4 and 5 the $R^2$ of the indicators is higher than .5 except for R6 (figure 1), all parameters (standard error terms and correlations) are different from zero and statistically significant (p <.05). In the case of model 5, item R19 displays a standardized correlation with factor 1 of about .8, and explains 64% of its variance, which would indicate that the effect of the rest of factors on this item is insignificant. According to the results of the goodness of fit indexes, model 5, based on five oblique factors, is the one that displays a better fit. The standardized solution for this model appears in figure 1.

FIGURE 1

## 3. 2 IDENTIFICATION OF LECTURER’S CATEGORIES

BIRCH defined that the simplest structure that represented homogenous solutions was a three-cluster solution, according to the Bayesian Information Criterion or Schwartz's Bayesian Criterion. Figure 2 shows the profiles of each cluster that represent the mean scores of the different items. Cluster 1 is composed of 26.2% of lecturer-subject observations containing all the items with low scores, lower than five in almost all cases. Cluster 2 comprises 45.2% of observations with intermediate scores between six and seven, and cluster 3 is composed of 28.6% of observations with scores higher than seven for nearly all items. The standard deviations for each item are highest in cluster 3, which indicates a greater degree of variability in terms of the students' ratings. It can also be seen that the profiles of the three different groups are nearly parallel, which indicates that the underlying factors are correlated throughout the questionnaire and the quantitative differences among clusters are indeed weak.

FIGURE 2

As a temporary validity check, we repeated the cluster analysis using data from the year 2000-2001, and once again we found that the optimal structure was the three-cluster solution. According to these results, three types of categories of lecturers could be suggested in terms of mean scores and it would be possible to obtain a quality score to rate lecturers using a few representative items of the questionnaire as suggested by Marsh (1994).

### 3.3 CHARACTERIZATION OF LECTURER'S CATEGORIES

CHAID produced a tree with a single branch based on the department variable and 6 terminal nodes (table 7) which is interpreted as follows. Departments grouped in each node are statistically similar. The first node yielded a total of 365 lecturer-subject combinations corresponding to 6 departments. The 19.18% of these observations belonged to cluster 1, 36.71% belonged to cluster 3, and 44.11% belonged to cluster 2. It should be stressed that only 2 departments (node 6) were predominantly associated with cluster 1 which is related to lecturers with low mean scores in all items. This node has the lowest number of observations,

TABLE 7

with only 23 subject-lecturers. By contrast, departments at nodes 2, 3 and 4 correspond to observations basically assigned to cluster 3, associated with higher scores. This result indicates that there is some association between some descriptive variables and mean scores assigned by students, but it is not so evident in the case of low scores due to the few number of observations in node 6.

Following this analysis, we studied what would happen if we fix the data split using faculty and course as our predictor variables. In both cases we obtained a significant split with different branches in each case, as shown in table 7. Using academic year as our predictor variable, we detected that lecturers of optional subjects, which accounting for 66 academic credits for the degree, were ranked similar to lecturers of the first courses, as they are included in the same node, while the second node (fourth and fifth courses) accounts for applied subjects more related to the professional occupation than earlier ones.

These results confirm that different descriptive variables contained in the MD matrix, especially department, academic year and faculty, have a clear explanatory capacity and a hierarchical influence on the perceived quality of teaching.

## 4. Discussion

The results of CFA validate that a model with five correlated dimensions is the best solution for explaining the underlying structure of the SET. This result is important, because although student ratings are an essential source of data for quality improvement at universities, it is necessary to take into account additional aspects of evaluation and the possible correlation between them. So, the results of the surveys should be considered with caution (Feldman, 1979), and is recommendable to use the data of student surveys together with other sources (Cashin, 1983). BIRCH allowed the identification of three lecturer categories, so teaching quality of lecturers cannot be treated as a homogenous group. This differentiation was not associated with particular items of the SET questionnaire which is reflected in the parallel structure of cluster profile. CHAID was performed to evaluate which variables included in the MD matrix were most capable to split the MS database. Particular descriptors such as department, faculty, and academic year were found to provide a statistically significant classification.

The different statistical analyses carried out are not only applicable to SET questionnaires but also to the evaluation of other aspects of a lecturer's work, such as research or management. Our method is expected to serve as a practical guideline that, subject to possible modifications in the future, may become an essential tool used by universities for their continuous improvement programs.

## References

Batista, J.M., Coenders, G. (eds) Modelos de Ecuaciones Estructurales. Cuadernos de estadística, 6. La Muralla, Madrid (2000)

Batista, J.M., Coenders, G., Alonso, J.: Análisis Factorial Confirmatorio. Su utilidad en validación de cuestionarios relacionados con la salud. Medicina Clínica (Barcelona), **122 (1),** 21-27 (2004)

Bentler, P.M.: Comparative fit indexes in structural models. Psychological Bull. **107(2),** 238-246 (1990)

Bentler, P.M., Bonett, D.G.: Significance tests and goodness of fit in the analysis of covariance structures. Psychol. Bull. **88**, 588-606 (1980)

Berk, R.A.: Survey of 12 strategies to measure teaching effectiveness. International Journal of Teach. and Learning in High. Educ. **17**, 48-62 (2005)

Bollen, K. A. (ed.). Structural Equations with Latent Variables. John Wiley & Sons, New York (1989)

Bollen, K.A., Long, J.S. (eds.) Testing Structural Equation Models. Sage, Newbury Park, California (1993)

Brown, M.W., Cudeck, R.: Alternative ways of assessing model fit. In: Bollen, K.A., Long, J.S. (eds.) Testing Structural Equation Models, pp. 136-162. Sage, Newbury Park, California (1993)

Cajide, J.: Análisis factorial confirmatorio de las características de calidad docente universitaria (solución LISREL). Bordón **46(4)**, 389-405 (1994)

Campbell, D.T., Fiske, D.W. Convergent and discriminant validation by the multitrait–multimethod matrix. Psychol. Bull. **56(2)**, 81–105 (1959)

Cashin, W.E.: Concerns about using student ratings in community colleges. New Dir. for Community Coll. **11(1)**, 57-65 (1983)

Chiu, T., Fang, D., Chen, J., Wang, Y., Jeris, C.: A robust and scalable clustering algorithm for mixed type attributes in large database environment. In: Proceedings of Seventh ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining, San Francisco, 2001

De la Orden, A.: Calidad y evaluación de la enseñanza universitaria. Congreso Internacional de las Universidades, Madrid, 1992

Feldman, K.A.: The significance of circumstances for college student's ratings of their teachers and courses. Res. in High. Educ. **10**, 149-172 (1979)

Göb, R.; McCollin, C., Fernanda Ramalhoto, M.: Ordinal methodology in the analysis of likert scales. Qualit. &Quantit. **41**, 601-626 (2007)

González Such, J. Estudio de un Instrumento para la Evaluación del Profesorado Universitario. Thesis. Universitat de València. Valencia (1997)

Gursoy D., Umbreit, W. T.: Exploring student´s evaluations of teaching effectiveness: what factors are important? Journal of Hosp. & Tour. Res. **29 (1)**, 91-109 (2005)

Hair, J. F., Anderson, R. E., Tatham, R. L., Black, W. C. (eds.) Multivariate data analysis. New York, Prentice Hall International (1995)

Hu, L., Bentler, P.M.: Cut off criteria fir fit indexes in covariance. Structure analysis: Conventional criteria versus new alternatives. Struct. Equ. Model. **6(1)**, 1-55 (1999)

Jackson, D.; Teal, C.R., Raines, J.S., Nansel, T.R., Force, R.C., Burdsal, C.A.: The dimensions of students’ perceptions of teaching effectiveness. Educ. and Psychol. Meas. **59(4)**, 580-596 (1999)

Jöreskog, K.G.: Testing structural equation models. In: K.A. Bollen & J.S. Long (eds.) Testing Structural Equation Models, pp. 294-316. Sage, Newbury Park (1993)

Jöreskog, K.G., Sörbom, D. (eds.) LISREL VI. Analysis of Linear Structural Relationships by Maximum Likelihood, Instrumental Variable, and Least Squares Methods. Scientific Software. Moresville, Indiana. (1986)

Jöreskog, K. G., Sörbom, D. (eds.) LISREl V. Analysis of Linear Structural Relations by the Method of Maximum Likelihood. SPSS Publications. Chicago, Illinois (1981)

Kass, G.V.: An exploratory technique for investigating large quantities of categorical data. Appl. Stat. **29(2)**, 119-127 (1980)

Lalla, M., Facchinetti, G., Maestroleo, G.: Ordinal scales and fuzzy set systems to measure agreement: An application to the evaluation of teaching activity. Qualit. & Quantit. **38**, 577-601 (2004)

Marsh, H. W.: Validity of students` evaluations of college teaching: a multitrait multimethod analysis. Journal of Educ. Psychol. **74(2)**, 219-237 (1982)

Marsh, H. W.: Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases, and utility. Journal of Educ. Psychol. **76(5)**, 707-754 (1984)

Marsh, H. W.: Students' evaluations of university teaching: Research, findings, methodological issues and directions for future research. Journal of Educ. Res. **11(3)**, 253 -388 (1987)

Marsh, H. W.: Weighting for the right criteria in the Instructional Development and Effectiveness Assessment (IDEA) system: Global and specific ratings of teaching effectiveness and their relation to course objectives. Journal of Educ Psychol. **86**, 631-648 (1994)

Marsh, H. W., Hattie, J.: The relation between research productivity and teaching effectiveness. Journal of High. Educ. **73**, 603-641 (2002)

Mulaik, Stanley A., James Larry R., Van Alstine, J., Bennett N., Lind S., Stilwell, C. D.: Evaluation of goodness-of-fit indices for structural equation models. Psychol. Bull. **105(3)**, 430-445 (1989)

Rodríguez Sabiote, C., Gutiérrez Pérez, J.: Debilidades de la evaluación de la calidad en la universidad española. Causas, consecuencias y propuestas de mejora. Revista Electron. de Investig. Educ. **5(1)** (2003) http://redie.ens.uabc.mx/vol5no1/contenido-sabiote.html. Retrieved, 20 April 2005.

Ting, K. F.: A multilevel perspective on student ratings of instruction: Lessons from the Chinese experience. Res. in High. Educ. **41**, 637-661 (2000)

Toland, M. D., De Ayala, R.J.: A multilevel factor analysis of student´s evaluations of teachig. Educ. and Psychol. Meas., **65(2)**, 272-296 (2005)

## Tables

*Table 1. Items of the SET (student evaluation of teaching) questionnaire used at UPV*

R1. The instructor communicates his/her ideas in a clear and ordered manner.
R2. It seems that he/she prepares the classes properly.
R3. The development of classes allowed following the explanations.
R4. The instructor remarks the important concepts in an objective and informative manner.
R5. The instructor answers the questions with clarity and interest.
R6. The subject matter was developed properly in the course.
R7. He/she teaches the subject with an applied focus related to competencies and points of view needed by professionals.
R8. He/she establishes connections with the contents of other subjects.
R9. The instructor encourages students to ask questions about the material.
R10. He / she talks to students about the progress of the classes and takes their opinions into consideration.
R11. The instructor was approachable and concerned about the progress of students.
R12. The instructor achieved students to be motivated by the subject.
R13. Examinations were representative of the material presented in the course.
R14. Examinations are corrected in a fair manner.
R15. Students have the possibility to review the exam and to comment all disagreements with the instructors.
R16. Recommended materials were useful and convenient for understanding the subject.
R17. The instructor makes an effective use of teaching aids (chalkboard, overhead projectors, slides, etc.).
R18.Theoretical contents were applied properly in exercises, assignments or practices.
R19. Taking into account the limitations, I think that the instructor that teaches the subject is a good lecturer.

*Table 2. Structure of IS and MS datasets*

| Academic year | Enquiries (IS) | Subject-lecturer combinations (MS) |
|---|---|---|
| 1995/1996 | 117702 | 3054 |
| 1996/1997 | 130756 | 3878 |
| 1997/1998 | 135582 | 4384 |
| 1998/1999 | 140149 | 4870 |
| 1999/2000 | 136267 | 5048 |
| 2000/2001 | 134314 | 5760 |

*Table 3. Structure Matrix*

| | Component[a] | | | | |
|---|---|---|---|---|---|
| Items | 1 | 2 | 3 | 4 | 5 |
| R1 | .916 | .459 | .558 | .602 | .494 |
| R2 | .881 | .463 | .523 | .628 | .475 |
| R3 | .835 | .503 | .546 | .525 | .548 |
| R4 | .838 | .545 | .589 | .579 | .603 |
| R5 | .828 | .475 | .649 | .593 | .528 |
| R6 | .755 | .642 | .482 | .605 | .635 |
| R7 | .637 | .449 | .533 | .599 | .862 |
| R8 | .578 | .428 | .603 | .531 | .885 |
| R9 | .656 | .488 | .843 | .584 | .598 |
| R10 | .613 | .510 | .902 | .552 | .556 |
| R11 | .676 | .551 | .893 | .585 | .491 |
| R12 | .752 | .567 | .773 | .606 | .632 |
| R13 | .550 | .885 | .413 | .549 | .440 |
| R14 | .517 | .883 | .535 | .517 | .389 |
| R15 | .470 | .790 | .536 | .629 | .265 |
| R16 | .613 | .560 | .510 | .893 | .485 |
| R17 | .642 | .510 | .541 | .902 | .519 |
| R18 | .706 | 651 | .538 | .770 | .532 |
| R19 | .855 | .610 | .711 | .694 | .562 |

[a] The highest value for each item indicates the most significant correlation on the specific factor.

*Table 4. Indexes of internal consistency and reliability for different models*[a]

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Cronbach's Alpha | .909 | .909 | .909 | .909 | .909 |
| Coefficient Alpha for an optimal short scale[b] | .941 | .941 | .941 | .941 | .941 |
| Reliability Coefficient RHO | .904 | .948 | .945 | .957 | .958 |
| Greatest Lower Bound (GLB) Reliability | .966 | .966 | | .966 | .966 |
| GLB Reliability for an optimal short scale[c] | .974 | .974 | | .974 | .974 |
| Bentler's dimension-free Lower Bound Reliability | .966 | .966 | | .966 | .966 |

[a]Model analyzed correspond to: model 1 hypothesized with one factor, model 2 hypothesized with two uncorrelated factors, model 3 hypothesized with two correlated factors; model 4 hypothesized with five correlated factors and all them related with R19 and model 5 hypothesized with five correlated factors; [b]Based on 16 variables, all except: R13, R14 and R15; [c] Based on three variables: R13, R14 and R15.

*Table 5. Construct reliability (CR) and variance extracted (VE)*

| | 5-Factor model[a] | | 2- Factor model | | 1-Factor model | |
|---|---|---|---|---|---|---|
| | CR[b] | EV[b] | CR | EV | CR | EV |
| Factor 1 | 0.8991 | 0.5182 | 0.9419 | 0.1773 | 0.9260 | 0.1384 |
| Factor 2 | 0.9742 | 0.9263 | 0.9742 | 0.9263 | | |
| Factor 3 | 0.7276 | 0.5719 | | | | |
| Factor 4 | 0.8773 | 0.5713 | | | | |
| Factor 5 | 0.8373 | 0.6320 | | | | |

[a]*Factor 1: Command, organization, and clarity of subject and program; Factor 2: Evaluation; Factor 3: Lecturer-student interrelation; Factor 4: Resources ; Factor 5: Interrelation with other subjects;* [b]Appropriate values for reliability and extracted variance, are higher than .7 and .5 respectively.

*Table 6. Goodness of fit indices for different models*[a]

| a | $\chi^2$ | df[b] | NFI[c] | NNFI[d] | CFI[e] | PNFI[f] | PGFI[g] | RMSEA[h] | IFI[i] | MFI[j] | GFI[k] | AGFI[l] | HOELTER | ECVI[m] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. | 754262.4[*] | 151 | .61 | .56 | .61 | .54 | .54 | .192 | .61 | .062 | .684 | .602 | 33 | 5.56 |
| 2. | 190316.5[*] | 152 | .90 | .89 | .90 | .80 | .80 | .096 | .90 | .496 | .841 | .801 | 130 | 1.40 |
| 3 | 189152.9[*] | 151 | .90 | .89 | .90 | .79 | .79 | .096 | .90 | .498 | .841 | .801 | 130 | 1.39 |
| 4. | 80254.2[*] | 139 | .96 | .95 | .96 | .80 | .80 | .065 | .96 | .744 | .942 | .920 | 386 | 0.44 |
| 5. | 43683.4[*] | 135 | .98 | .97 | .98 | .79 | .79 | .049 | .98 | .852 | .966 | .952 | 517 | 0.32 |

[a]Models analyzed as described in table 4; [b]df =Degree of Freedom; [c]NFI = Bentler-Bonett Normed Fit Index; [d]NNFI = Bentler-Bonett Not-Normed Fit Index; [e]CFI = Comparative Fit Index; [f]PNFI = Parsimony Normed Fit Index; [g]PGFI = Parsimonious Goodness of Fit Index; [h]RMSEA = Root Mean Squared Error of Approximation; [i]IFI = Bollen Fit Index; [j]MFI = McDonal fit index; [k]GFI = Goodness-of-Fit Index; [l]AGFI = Adjusted Goodness-of-Fit Index; [m]ECVI = Index of Cross validation; [*]p < .05.

*Table 7. Results of CHAID*

| Predictor[a] | Node[b] | Cluster[c] | % Observations[d] | Second Branch[e] | Nn2b[f] |
|---|---|---|---|---|---|
| Department[g] | 1 | 1 | 19.18 | | |
| | | 2 | 44.11 | | |
| | | 3 | 36.71 | | |
| | 2 | 1 | 16.21 | | |
| | | 2 | 37.59 | | |
| | | 3 | 46.21 | | |
| | 3 | 1 | 24 | | |
| | | 2 | 28.73 | | |
| | | 3 | 47.27 | | |
| | 4 | 1 | 3.61 | | |
| | | 2 | 45.78 | | |
| | | 3 | 50.60 | | |
| | 5 | 1 | 19.57 | | |
| | | 2 | 56.52 | | |
| | | 3 | 23.91 | | |
| | 6 | 1 | 43.48 | | |
| | | 2 | 34.78 | | |
| | | 3 | 21.74 | | |
| Faculty[h] | 1 | 1 | 7.14 | Department | 3 |
| | | 2 | 50 | | |
| | | 3 | 42.66 | | |
| | 2 | 1 | 15.09 | | |
| | | 2 | 26.42 | | |
| | | 3 | 58.49 | | |
| | 3 | 1 | 16.89 | | |
| | | 2 | 41.63 | | |
| | | 3 | 41.48 | | |
| | 4 | 1 | 28.31 | Department | 2 |
| | | 2 | 36.14 | | |
| | | 3 | 35.54 | | |
| Academic Year[h] | 1 (1st, 2nd, 3rd, 6th and Optional Subjects) | 1 | 17.13 | Department | 6 |
| | | 2 | 40.81 | | |
| | | 3 | 42.07 | | |
| | 2 (4th and 5th) | 1 | 22.90 | Department | 4 |
| | | 2 | 39.21 | | |
| | | 3 | 37.89 | | |

Note:

[a]Descriptive variables of MD dataset; [b]Node number of the first split obtained with the corresponding predictor; [c]Clusters obtained with BIRCH; [d]Percentage of observations for each node classified in each cluster; [e]Second Branch obtained with the second predictor of the splitting; [f]Nn2b: Number of nodes in second branch; [g]Split predictor variable automatically selected by CHAID; [h]Other descriptive variables such as faculty and academic year that were fixed to split the dataset.

**Figure**

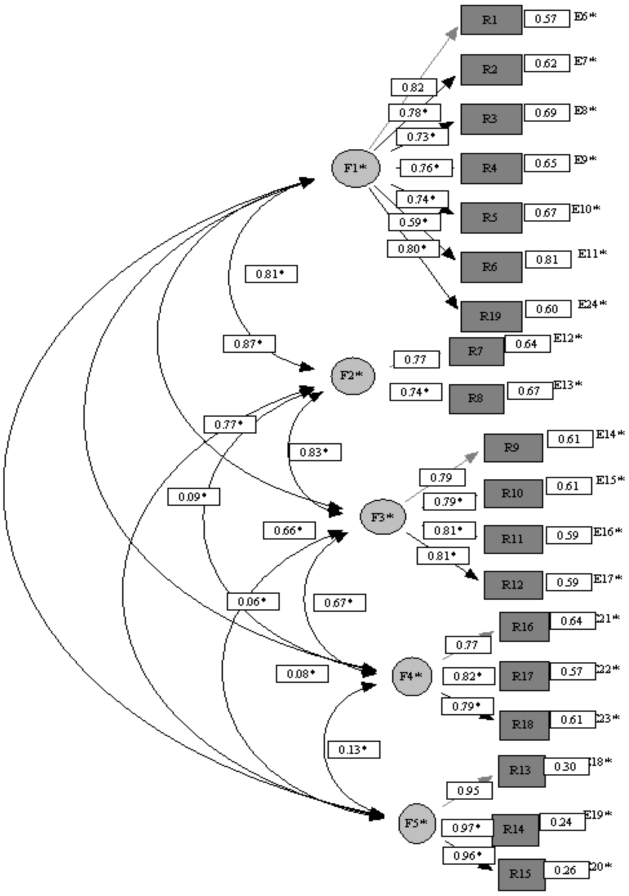

*Fig. 1. Standardized solution for model 5.* Standardized regression coefficients (the asterisk indicates p<.05) and the terms of measurement error *(1-$R^2$)* of each item or indicator. $R_i$: item or observable variable; $E_i$: error term; Latent variables are enclosed in ellipses; Double-ended arrows: relationships between latent variables; We fixed to one the initial saturation coefficients corresponding to a single variable for each of the factors to avoid identification problems.

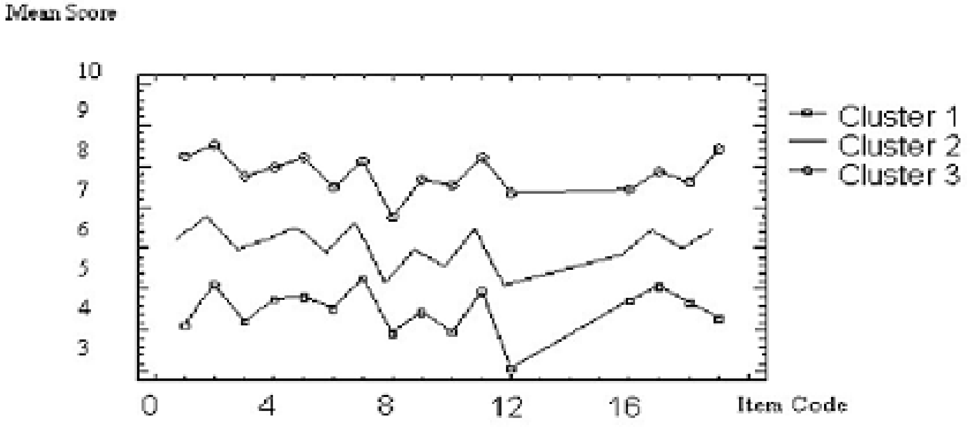

*Fig. 2. Cluster profiles values obtained with BIRCH for the year 1999/2000.* Cluster 1 is comprised by observations (i.e. lecturer-subjects) containing all items with scores lower than five in almost all cases. Cluster 2 accounts for observations with scores between six and seven. Cluster 3 is composed of observations with scores higher than 7 for almost all items. Items R13, R14 and R15 were not included.