Developments in Statistics
Andrej Mrvar and Anuška Ferligoj (Editors)
Metodološki zvezki, 17, Ljubljana: FDV, 2002

# A Statistical Analysis of Teaching Effectiveness from Students' Point of View

Laura Pagani[1] and Chiara Seghieri[2]

**Abstract**

Teaching is a multidimensional process comprising a number of aspects, e.g., instructor attributes, which sometimes are difficult to evaluate. In particular teaching effectiveness, that is an aspect of teaching, is influenced by a combination of teacher characteristics (such as clarity, capacity to motivate the students and to help them in the study of his topic, ability to organize the lesson also with exercises and handouts, for example but also gender, age, previous experiences), physical aspects of the classroom or laboratory (too crowded or with an insufficient number of computers) and class characteristics (such as students' characteristics: gender, age, high-school of origin, mark obtained at the end of compulsory or high school, faculty attended by the student, or class size). As teaching effectiveness is becoming even more important in a system of school evaluation (in our case university system of evaluation), it is necessary to find how to measure it. This paper considers the problem of assessment of teaching effectiveness from students' point of view, analysing the questionnaires given to the students of the University of Udine at the end of their courses. The problem, in statistical terms is to relate an "outcome" variable (the dependent one), in this case ratings given by the students or a particular linear combination of it, to a set of explanatory variables both to the student and teacher level. The data set used in the analysis consists of almost 9500 questionnaires regarding 416 courses of the University of Udine covering the academic year 1999-2000, its structure (questionnaires clustered in courses) suggest the use of a particular class of regression models, the multilevel models.

[1] Department of Statistics, University of Udine, Via Treppo, 18, 33100 Udine, Italy; pagani@dss.uniud.it.

[2] Department of Statistics "G. Parenti", University of Florence, Viale Morgagni, 59, 50134 Florence, Italy; seghieri@ds.unifi.it.

# 1 Introduction

In the last years an important problem, from social and political point of view, is the assessment of public sector activities (education, health, social services) with the aim to compare institutions or subjects. In the field of education, if the focus is the evaluation of "effectiveness", with the intent on comparing school or teachers, the use of "outcome" indicators such as examination results or final grades.

The principal aim of this paper is to evaluate the teaching effectiveness of a sample of instructors at the University of Udine using data coming from a questionnaire given to the students at the end of each academic course.

As teaching effectiveness is becoming even more important in a system of school's evaluation (in our case university system of evaluation), it is necessary to find a measure of it.

Teaching is a multidimensional process comprising a number of separable dimensions or instructor attributes, which sometimes are difficult to evaluate in a quantitative way (Arreola, 1995; Centra, 1993; Boex, 2000). An instructor's overall reaching effectiveness, that is an aspect of teaching, is influenced by a combination of teacher characteristics (such as clarity, capacity to motivate the students and to help them in the study of his topic, ability to organize the lesson also with exercises and handouts, for example but also gender, age, previous experiences), physical aspects of the classroom or laboratory (too crowded or with an insufficient number of computers) and class characteristics (such as students' characteristics: gender, age, high-school of origin, mark obtained at the end of compulsory or high school, faculty attended by the student, or class size).

On the issue surrounding the study of the students' evaluations to measure the teaching effectiveness there is a debate.

Proponents of the multidimensional view of education process argue that, because of the multidimensional nature of teaching, instruction can not be capture by on single measure such as a global effectiveness rating (Marsh, 1987). Using factor analysis Marsh identified nine separate dimensions of teaching (learning, enthusiasm, organization, group interaction, individual rapport, breadth of coverage, examination/grading, assignments and workload/difficulty). He concluded that each of these dimensions is important and each of them has to be exanimate to evaluate the instructors.

However Abrami (1989) recognized that the nature of effective teaching could vary across instructors, courses, students and settings. He recommended using global evaluation items whenever summative judgements about teaching effectiveness are called for.

A compromise between these positions has been suggested calling for the items of the questionnaire to be weighted to evaluate an overall measure (Ryan and Harrison, 1995).

In literature we found studies employing regression techniques to study the relationships between the indicator of effectiveness and courses and students' characteristics.

In particular we conduct the analysis in two steps:

1. We measure the instructor's effectiveness adopting two different kind of indicators: the first obtained from the last global question of the questionnaire on the level of course/teacher satisfaction and the second the first component obtained from a principal component analysis (PCA) performed on the 18 items of the questionnaire (Dillon and Goldstein, 1984).
2. We fit two kinds of multilevel models using as response variables the indicators mentioned at point 1.

There is a considerable literature on multilevel models used to evaluate education, or more in general, public sector activities, see for example Aitkin and Longford (1986), Goldstein and Spiegelhalter (1996), Goldstein (1997), but it seems that no previous studies attempt to evaluate teachers' effectiveness using these techniques.

In the next section we describe the data set, then in section 3 the statistical analysis with the models we use and the results are presented together with goodness of fit measures. In the final section (section 4) we draw some conclusions and discuss potential development.

# 2 Data

To study the effectiveness of instruction we use data from a questionnaire given to the students at the end of each course.

The questionnaire is divided in two parts: the first collects information on the students' characteristics (age, gender, type of high school attended and so on); the second consists of 18 general items about teacher characteristics (instructor's teaching qualities, materials adopted) and a last global item on the level of course-instructor satisfaction. Response is measured on a five-points scale ranging from 1 (not at all satisfactory) to 5 (very satisfactory).

The data set used in this study consists of 9561 questionnaires regarding 416 courses of eight faculties (Agriculture, Engineering, Medicine, Letters, Languages, Economy, Science, Veterinary medicine) of the University of Udine covering the academic year 1999-2000.

The structure of the data set is quite complex because there is not a one-to-one correspondence between questionnaire and student and between course and teacher. Following Rasbash and Browne (2001) we can identify two crossed hierarchies in the data: instructors and courses (an instructors teaches in more than one course) and students and questionnaires (a student compiles more than one questionnaire). The situation of the data structure can be represented in the following diagram.

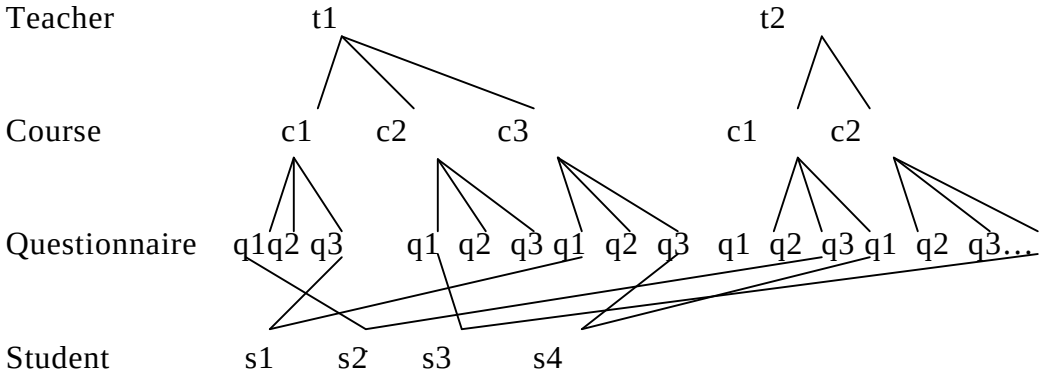


**Figure 1:** Diagram for a data structure with more than one cross-classified structure.

In Figure 1 the questionnaires are positioned within courses within instructors, so the diagram reflects the hierarchy for teachers. When we connect the students hierarchy to the diagram we can see the cross classification between students and questionnaires, highlighting the crossed structure of data set. But obviously the questionnaires are anonymous, moreover we have only the course code and not the name of the instructor, so, without any additional information on the data, we have to consider them as hierarchical and use basic multilevel models also if there are negative consequences in ignoring the non hierarchical structure of the data set (under-specification of the model because that do not include sources of variation, variance components that can not be trusted). Another important concern with regard to the data set is the possible presence of selection bias due to the fact that course attendance is not compulsory and to the fact that questionnaires are distributed during the last days of course so that many students may not participate in the evaluation.

# 3 Multilevel models for point scores

The main objective is to use the multilevel analysis to relate an “outcome” variable (the dependent one), in this case ratings given by the students or a particular linear combination of it, to a set of explanatory variables both to the student and teacher level.

As we pointed out in the previous section we consider the data as hierarchical with two levels: the level-1 units ($i$=1,…,9561) are the questionnaires, while the level-2 units ($j$=1,…,416) are courses.

We used two different kinds of response variable and three different multilevel models.

In the first multilevel model (model A), we consider as a response variable the last global item, that is ordinal, as linear. Assuming that the rating are converted in arbitrary point scores, ranging from 1 (not at all satisfactory) to 5 (very satisfactory) we consider the ordered response variable as a continuous normal one. With these assumptions about the response variable the hypotheses of the multilevel model are:

$$_{A}Y_{ij} = \alpha_0 + \sum_{h=1}^{H} \alpha_h x_{hij} + \sum_{k=1}^{K} \beta_k z_{kj} + U_{0j} + R_{ij}, \quad U_{0j} \sim \mathrm{N}(0, \sigma_u^2), \; R_{ij} \sim N(0, \sigma_e^2)$$

In the second multilevel model (model B), an ordered probit one, the response variable is the last global item of the questionnaire. The response unobservable variable, say "level of satisfaction", is denoted by $_{B}Y^*$ . Following Snijders and Bosker (1999) the random intercept ordered category model, with $H$ explanatory variables, for $_{B}Y^*$ is

$$_{B}Y_{ij}^{*} = \gamma_o + \sum_{h=1}^{H} \gamma_h x_{hij} + \sum_{k=1}^{K} \lambda_k z_{kj} + U_{0j} + R_{ij}, \quad R_{ij} \sim N(0,1)$$

The observed response variable, denoted by $_{B}Y$ , is related to $_{B}Y^*$ as follows

$$_{B}Y = \begin{cases} 1 & if \; -\infty < {}_{B}Y^* \le \theta_1 \\ 2 & if \; \theta_1 < {}_{B}Y^* \le \theta_2 \\ 3 & if \; \theta_2 < {}_{B}Y^* \le \theta_3 \\ 4 & if \; \theta_3 < {}_{B}Y^* \le \theta_4 \\ 5 & if \; \theta_4 < {}_{B}Y^* < +\infty \end{cases}$$

where $\theta_1, \theta_2, \theta_3, \theta_4$ are the thresholds parameters. In the third multilevel model (model C) the response variable is the first component obtained from a principal component analysis (PCA) performed on the 18 items of the questionnaire (Dillon and Goldstein, 1984). This component accounted for 45.5 percent of the total variation. We consider this variable as a continuous and normal one. The main results of this analysis are summarised in Table 1.

As the weights are quite similar we can consider the first component as an "average" of the 18 items. So the response variable for this model, denoted by $_{C}Y$, is the "average level of satisfaction". The hypotheses of this model are:

$$_{C}Y_{ij} = \alpha_0 + \sum_{h=1}^{H} \alpha_h x_{hij} + \sum_{k=1}^{K} \beta_k z_{kj} + U_{0j} + R_{ij}, \quad U_{0j} \quad \mathrm{N}(0, \sigma_u^2), \; R_{ij} \sim \; N(0, \sigma_e^2)$$

All the response variables we adopted can be seen as a synthesis of the level of students' satisfaction. While the first and the second are measures of satisfaction obtained from one global item, the third is a weighted sum of the information contained in the 18 original items and so it captures multidimensional view of teaching attributes.

**Table 1:** PCA analysis on the 18 items – first component.

| | Questionnaire Item | Weights | Correlation |
|---|---|---|---|
| The Instructor | Meets course objectives | 0.281 | 0.804 |
| | Indicates how to prepare the course | 0.271 | 0.778 |
| | Develops the course sistematically | 0.278 | 0.797 |
| | Outlines the major points clearly | 0.288 | 0.825 |
| | Links to other subjects | 0.227 | 0.651 |
| | Provides examples and case studies | 0.266 | 0.761 |
| | Explains clearly | 0.281 | 0.804 |
| | Motivates the students | 0.278 | 0.797 |
| | Gives deeper understanding of the concepts | 0.248 | 0.712 |
| | Is punctual | 0.227 | 0.651 |
| | Is accessible to students out of class | 0.221 | 0.631 |
| | Has a genuine interest in students | 0.255 | 0.731 |
| Teaching Material | Quality of text books and teacher notes | 0.194 | 0.558 |
| | Effectiveness of other teaching materials | 0.209 | 0.601 |
| | Quantity of time dedicated to practiceand exercises | 0.153 | 0.439 |
| | Utility of exercises, laboratory exercises,… | 0.153 | 0.439 |
| | Coordination between lectures and exercises | 0.161 | 0.461 |
| | Satisfation level of practices and exercises | 0.157 | 0.451 |

*Proportion of explained variance* : 45.5

# 4 Comparison of results between the three models

In this section we compare the parameter estimates for the three models: Model A, Model B and Model C. We tried to fit several models with different sets of explanatory variables using *MlwiN* (Goldstein et al., 1998), a standard software for multilevel models, for model A and model C and GLLAMM program for Stata (Rabe-Hesketh et al., 2001) for model B. The definitions of the explanatory variables used in the final models are reported in Table 2.

Table 3 provides the comparative results for models with only the intercept or threshold parameters (empty models). The results about the final models are shown in Table 4.

**Table 2:** Explanatory variables definitions.

| Variable Name | Definition |
|---|---|
| Dimension | The number of questionnaires for each course; it is an approximated measure of the class size |
| Not compulsory | 1= the course was a non compulsory one; 0= otherwise |
| Letters | 1= the student attended the faculty of Letters; 0= otherwise |
| Languages | 1= the student attended the faculty of Languages; 0= otherwise |
| Engineering | 1= the student attended the faculty of Engineering; 0= otherwise |
| Economy | 1= the student attended the faculty of Economy; 0= otherwise |
| Very good | 1= the student got "very good" as final grades for the first three years of secondary school (children from 11 to 14 years of age); 0= otherwise |
| Excellent | 1= the student got "excellent" as final grades for the first three years of secondary school (children from 11 to 14 years of age); 0= otherwise |
| Magistrali | 1= the student attended a high school for the training of primary teachers; 0= otherwise |
| Liceo classico | 1= the student attended a high school specializing in classics subjects; 0= otherwise |
| Lessons | 1= the student attended the most part of the course (over the 60%); 0=otherwise |
| Grade | Is the standardized mark obtained at the end of the high school. |
| Regularity | 1= the student attended the course at the institutional year of his academic career; 0=otherwise |

As shown in Tables 3 and 4 the estimate of the random parameters are significant for the three models. This means that the variability in the level of satisfaction depends on differences among courses.

If we compare the results in Table 3 with those in Table 4 we note that the introduction of fixed effects reduces the variances both at questionnaire and course level, this means that the introduction of individual and/or contextual variables reduces the unexplained variability at the two levels.

Following Snijders and Bosker (1999) we can estimate the level-one proportion of explained variance of the three models using the index

$$R_M^2 = 1 - \frac{\hat{\sigma}_{e,M}^2 + \hat{\sigma}_{u,M}^2}{\hat{\sigma}_{e,E}^2 + \hat{\sigma}_{u,E}^2}$$

where $\hat{\sigma}_{e,M}^2 + \hat{\sigma}_{u,M}^2$ is the estimated residual variance (or mean squared prediction error) for model M and $\hat{\sigma}_{e,E}^2 + \hat{\sigma}_{u,E}^2$ is the estimated residual variance (or mean squared prediction error) for the empty model. The results for Model A,

Model B and Model C are $R_A^2 = 0.13$, $R_B^2 = 0.14$ and $R_C^2 = 0.14$ and they show that these explained variances are quite low.

**Table 3.** Parameters estimation for model A, model B and model C (empty models).

| Parameter | Coefficient and S.E. model A | Coefficient and S.E. model B | Coefficient and S.E. model C |
|---|---|---|---|
| $\theta_1$ | | | |
| $\theta_2$ | | | |
| $\theta_3$ | | -2.304 (0.035) | |
| $\theta_4$ | | -1.314 (0.024) | |
| | | 0.354 (0.022) | |
| | | 1.670 (0.026) | |
| Intercept | 3.272 (0.025) | | 0.062 (0.089) |
| $\sigma_u^2$ | 0.157 (0.016) | 0.413 (0.023) | 2.095 (0.201) |
| $\sigma_e^2$ | 0.530 (0.009) | 1 | 5.070 (0.092) |
| *-2*loglikelihood* | 16895.52 | 21933.15 | 29449.75 |

**Table 4:** Parameters estimation for model A, model B and model C (final models).

| Parameter | Coefficient and S.E. model A | Coefficient and S.E. model B | Coefficient and S.E. model C |
|---|---|---|---|
| $\theta_1$ | | -2.478 (0.414) | |
| $\theta_2$ | | -1.483 (0.032) | |
| $\theta_3$ | | 0.193 (0.030) | |
| $\theta_4$ | | 1.517 (0.34) | |
| Intercept | 3.208 (0.039) | | 0.071 (0.164) |
| Dimension | - | -0.004 (0.0002) | - |
| Not compulsory | 0.195 (0.043) | - | 0.792 (0.148) |
| Letters | - | 0.601 (0.111) | - |
| Languages | - | 0.296 (0.060) | - |
| Engineering | -0.219 (0.061) | -0.130 (0.038) | -1.070 (0.214) |
| Economy | -0.231 (0.068) | - | -0.909 (0.236) |
| Very good | - | - | 0.128 ( 0.07) |
| Excellent | 0.034 (0.020) | - | 0.219 (0.078) |
| Magistrali | - | - | 0.334 (0.164) |
| Liceo classico | - | - | 0.181 (0.111) |
| Lessons | - | - | 0.169 (0.092) |
| Grade | - | - | 0.084 (0.032) |
| Regularity | - | - | -0.261 (0.097) |
| $\sigma_u^2$ | 0.136 (0.014) | 0.327 (0.016) | 1.726 (0.170) |
| $\sigma_e^2$ | 0.529 (0.009) | 1 | 5.034 (0.091) |
| *-2*loglikelihood* | 16849.65 | 21848.68 | 29354.53 |

Note: - indicates that the parameter is not statistically significant in that model

Analysing Table 4 we can point out that the results are different in term of covariate. In fact the only common result regards the Faculty of Engineering, with

a negative effect, this means that students attending this faculty give to the instructors lower rating. Other results are the negative effect on rating of the Faculty of Economy, and class size, and the positive effect of very good previous school experience and the kind of courses (compulsory or not), (for Model A and Model B).

Some general comments can be made about the results in Table 4. Firstly we didn't find significative interactions or more complex variance specifications (e.g., random slopes) in the three models. Secondly Model A and Model B seem to be insensitive to previous student school experience. Lastly, as the course effect is significant in all the three model, it is interesting to compare estimates of courses residuals, as they often are the basis for value-added (or effectiveness) indicators that are very important in educational effectiveness research. The result of this comparison is illustrated in Table 5 where the correlation coefficients, for the residuals and the rank correlations, for the ranks are calculated and in Figure 2 and Figure 3 .

**Table 5:** Correlation coefficients and rank correlations for model A, model B and model C (final models).

Course residuals

| | Model A | Model B | Model C |
|---|---|---|---|
| Model A | 1 | | |
| Model B | 0.883 | 1 | |
| Model C | 0.884 | 0.819 | 1 |

Ranks of course residuals

| | Model A | Model B | Model C |
|---|---|---|---|
| Model A | 1 | | |
| Model B | 0.868 | 1 | |
| Model C | 0.861 | 0.801 | 1 |

Note that the correlation coefficients and the rank correlations (and the scatter plots) suggest a strong agreement for the course value added estimates between the three models, especially for course with low and high value added.

This means that if an instructor (by means of his course) has a very low or very high effectiveness it is not important the way we use to measure the outcome variable, but this is not true for the others.

# 5 Conclusions

Our purpose in this study was to find which characteristics of students, instructors or courses influence teaching effectiveness and then to evaluate their effects on the response variables adopted.

The results are different in term of covariate in the three models adopted. In fact the only common result regards the Faculty of Engineering, with a negative effect, other results are the negative effect on rating of the Faculty of Economy, and class size, and the positive effect of very good previous school experience and the kind of courses (compulsory or not), (for Model A and Model B).

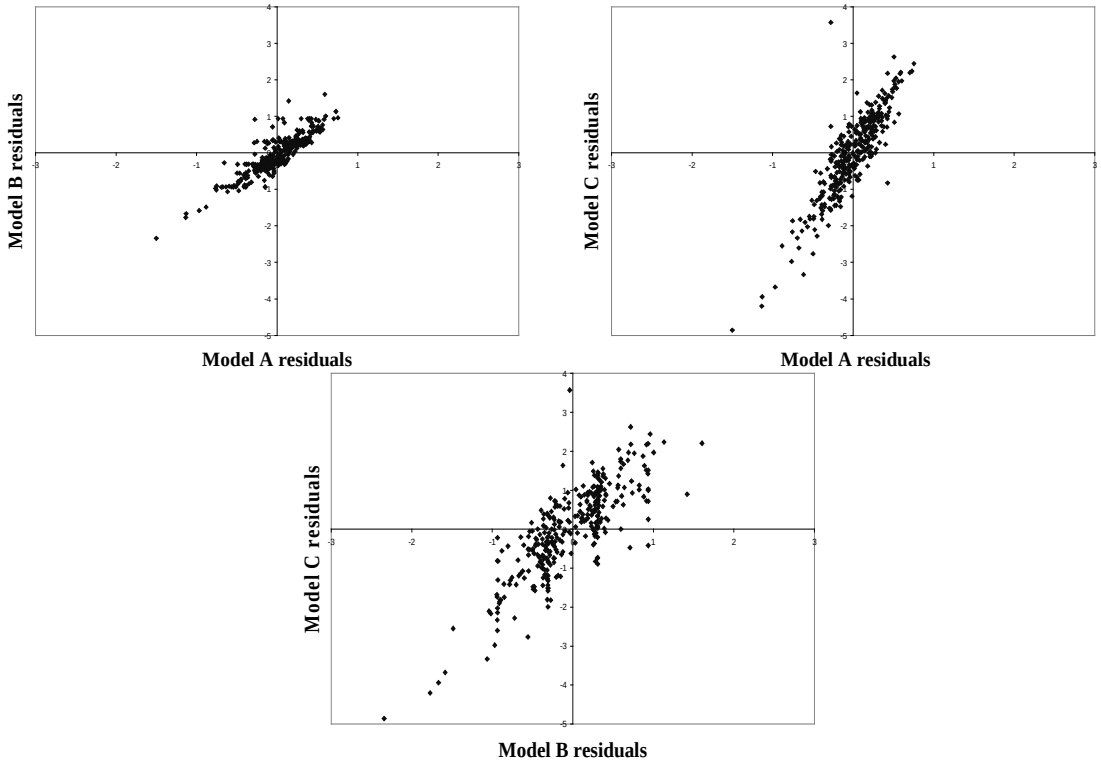


**Figure 2:** Plots of residuals of Model A against Model B, Model A against Model C and Model B against Model C.

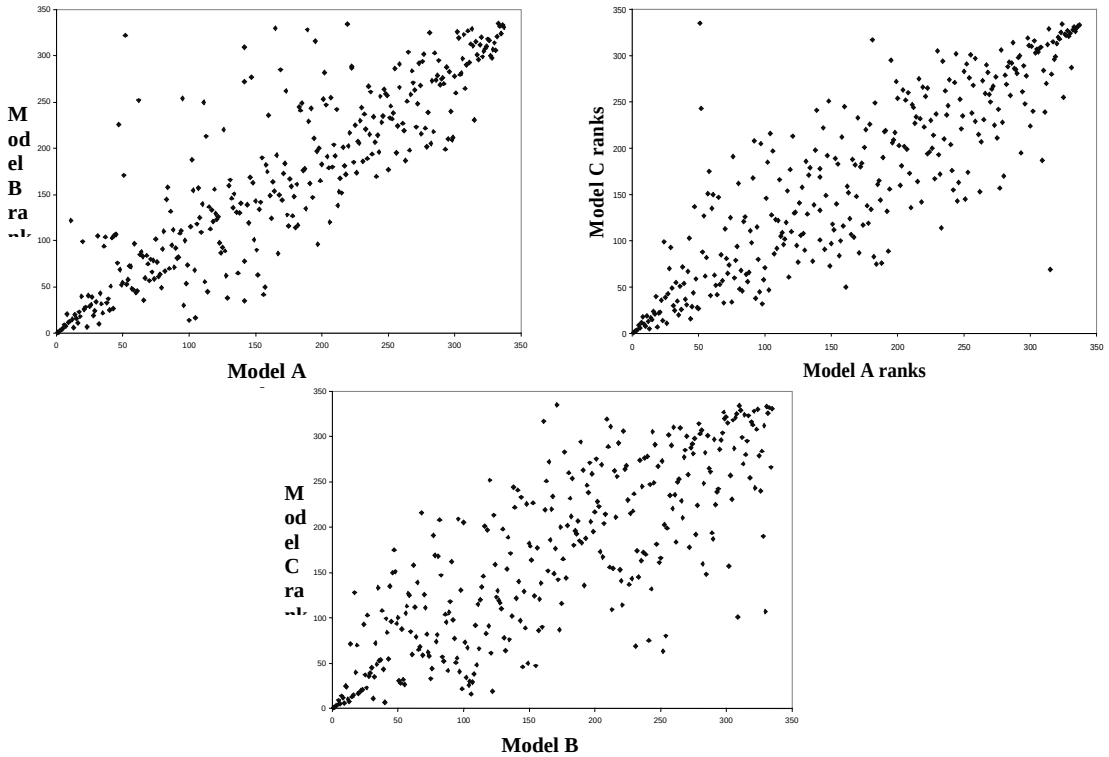


**Figure 3:** Plots of residual ranks of Model A against Model B, Model A against Model C and Model B against Model C.

So Model A and Model B seem to be insensitive to previous student school experience. As the course effect is significant in all the three model, it is interesting to compare estimates of courses residuals, as they often are the basis for value-added (or effectiveness) indicators that are very important in educational effectiveness research.

Analysing the course residuals and their ranks we note that if an instructor (by means of his course) has a very low or very high effectiveness it is not important the way we use to measure the outcome variable, but this is not true for the others.

There is another question arising from our study, due to the three different measures of teaching effectiveness we adopt (the global item, continuous and ordinal, and the first principal component).

In fact the three models (model A, Model B and Model C) provide different set of significative variables. Both models give reasonable results. We can't suggest a clear indication about how well one, as opposed to more specific items, reflects instructional effectiveness. The interpretation here is that both approaches are defensible.

Unless teaching is viewed conceptually as a single behaviour, there is theoretical justification for considering multiple items to assess the quality of teaching. Also, to the extent that evaluations are to be used formatively, specific measures are necessary to identify particular strengths and behaviours upon which individual instructors can improve.

Also, the extent that student ratings are one source (among many) of information for personnel decisions, the use of one global item is desirable on the basis of practicality.

So all the three models give useful indications to the instructors and institutions to improve the quality of teaching.

# References

[1] Abrami, P.C. (1989): How should we use student rating to evaluate teaching? *Research in Higher Education,* **30***,* 221-27.

[2] Aitkin, M. and Longford, N. (1986) : Statistical modelling issues in school effectiveness studies. *J. R. Stat. Soc. A*, **149**, 1-43.

[3] Arreola, R.A. (1995): *Developing a Comprehensive Faculty Evaluation System*. Bolton, Mass.: Anker.

[4] Boex, Jameson L.F. (2000) : Identifying the attributes of effective economics instructors: An analysis of student evaluation of instructor data. *Journal of Economic Education,* **31,** 211-26.

[5] Centra, J.A. (1993): *Reflective Faculty Evaluation*. San Francisco: Jossey-Bass.

[6] Dillon, W.R. and Goldstein, M. (1984): *Multivariate Analysis*. New York: Wiley.

[7] Goldstein, H. and Spiegelhalter, D. J. (1996): League tables and their limitations: Statistical issues in comparisons of institutional performance. *J. R. Stat. Soc. A*, **159**, 385-443.

[8] Goldstein, H. (1997): Methods in school effectiveness. *Research. School Effectiveness and School Improvement*, **8**, 369-395.

[9] Goldstein H. et al. (1998): *A User's Guide to Mlwin*. Institute of Education, University of London.

[10] Marsh, H.W. (1987): Student's evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, **11**, 263-388.

[11] Rabe-Hesketh, S. et al. (2001): *Gllamm Manual*. Technical report 2001/01, Department of Biostatistics and Computing, Institute of Psychiatry, King's College, University of London.

[12] Rasbash, J. and Browne, W.J. (2001): Non-hierarchical multilevel models. In A. Leyland and H. Goldstein (Eds): *Multilevel Modelling of Health Statistics*. New York: Wiley.

[13] Ryan, J. M. and Harrison, P.D. (1995): *The Relationship between Individual Instructional Characteristics and the Overall Assessment of Teaching Effectiveness across Different Instructional Contexts.* Research in Higher Education.

[14] Snijders, T. and Bosker, R. (1999): *Multilevel Analysis*. London: Sage Publications.