ISSN 0101-4161

Factors Affecting the Student Evaluation of Teaching Scores: Evidence from Panel Data Estimation♦

Eduardo de Carvalho Andrade
Professor Associado do Instituto de Ensino e Pesquisa (Inper)
Endereço para contato: Rua Quatá 300, 4º. andar - São Paulo - CEP: 04546-042
E-mail: eduardo.andrade@insper.edu.br

Bruno de Paula Rocha
Professor Adjunto da Universidade Federal de Minas Gerais (UFMG)
Centro de Desenvolvimento e Planejamento de Minas Gerais (CEDEPLAR)
Endereço para contato: Av. Antônio Carlos, 6627, Belo Horizonte - Brazil - CEP: 31270-901
E-mail: brunor@cedeplar.ufmg.br

Recebido em 11 de março de 2011. Aceito em 14 de dezembro de 2011.

Abstract
We use a random-effects model to find the factors that affect the student evaluation of teaching (SET) scores. Dataset covers 6 semesters, 496 undergraduate courses related to 101 instructors and 89 disciplines. Our empirical findings are: (i) the class size affects negatively the SET score; (ii) instructors with more experience are better evaluated, but these gains reduce over time; (iii) participating in training programs, designed to improve the quality of teaching, did not increase the SET scores; (iv) instructors seem to be able to marginally 'buy' a better evaluation by inflating students' grade. Finally, there are significant changes in the rankings when we adjust the SET score to eliminate the effects of variables beyond instructors' control. Despite these changes, they are not statistically significant.

Keywords
student evaluation score, random-effects model, undergraduate, ranking

JEL Classification
A22

Resumo
Este trabalho emprega um modelo de efeitos aleatórios para estimar os principais fatores determinantes na avaliação de professores por estudantes. Os dados compreendem

♦ We are grateful to Marcia Moura for authorizing the use of the data for this study. We would like to thank Carolina Costa, Tadeu Ponte and specially Rogério Costa for making available the data used for this study. Any views expressed are those of the authors' exclusively. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the Insper. We also would like to thank the three anonymous referees for helpful comments. Needless to say, remaining errors and omissions are of our responsibility.

496 cursos na graduação, relacionados a 101 diferentes professores e 89 disciplinas, durante 6 semestres letivos. Os principais resultados obtidos são: (i) o tamanho das salas de aula afeta negativamente a nota recebida pelo professor; (ii) professores mais experientes são mais bem avaliados, mas estes ganhos são decrescentes ao longo do tempo; (iii) a participação em programas de treinamento, desenhados para melhorar a qualidade do ensino, não afetam as notas recebidas pelos professores; (iv) a avaliação recebida pelos professores é marginalmente influenciada pela nota dada aos alunos durante o curso. Finalmente, os dados mostram mudanças significativas no ranking dos professores, quando são feitos ajustes para eliminar os efeitos de variáveis fora do controle dos professores. Entretanto, estas mudanças não são estatisticamente significativas.

Palavras-Chave
avaliação de professores, modelo de efeitos aleatórios, ensino de graduação, classificação de professores

1. Introduction

In several higher education institutions, it is common that students evaluate their professors in the end of the courses.[1] The results of the student evaluation of teaching (SET) are considered as an instrument to assess the quality of an instructor's teaching and are used by these institutions for purposes of promotion of the instructors.[2] Reflecting the importance of this topic for professors and universities' managers, there is a vast literature on the factors that affect the SET scores.[3]

A question of fundamental importance is if the instrument SET is appropriate for the purpose of assessing the quality of an instructor's teaching and, as a consequence, its use for deciding the instructor's promotion. In particular, the SET can be distorted in some undesirable ways. For example, an instructor may be able to 'buy' a better evaluation by inflating students' grades or the SET can be affected by variables beyond the instructor's control (such as age, sex or class size). When a higher education institution's manager does not take into consideration these possible effects, an instructor may be pro-

[1] For example, Becker and Watts (1999) show that this is the case for most departments of economics in the United States.
[2] Many studies have analyzed if higher SET scores in fact mean that the teaching quality is greater. The results are mixed. See, for example, Soper (1973) and Gramlich and Greenlee (1993).
[3] For a review of the literature, see McPherson *et al.* (2007).

moted or not unfairly. This paper's objective is to examine empirically such possible effects and to present an alternative to eliminate these possible distortions in order to maintain the usefulness of the SET for purposes of promotion.

This paper takes advantage of a new large panel data from Insper (a private higher education institution) with six semesters for the period from the second semester of 2005 to the first semester of 2008, encompassing 496 undergraduate courses taught by 101 instructors in 89 different disciplines. We use a random-effects model estimated with feasible generalized least squares to examine the effects of instructor-specific time-invariant characteristics as well as to control for unobservable characteristics of individual instructors. In these regards, the closest papers to this one in the literature are McPherson (2006) and McPherson *et al*. (2007). To our knowledge, this is the first time that such analysis is performed using data from a Brazilian higher education institution (HEI).

We find robust empirical evidence that some course's, instructor's and student's characteristics can affect the SET scores. The coefficients of the variables related to the instructor training programs though were not significant. Instructors seem to be able to marginally 'buy' a better evaluation by inflating student's grade. Other results found in the literature are also obtained here: the class size affects negatively the SET score and instructors with more experience are better evaluated, but these gains reduce over time.

Furthermore, we construct different instructors' rankings by adjusting the SET scores, in order to eliminate either the possible effects of instructors' manipulation through grade inflation or the effects of variables beyond the instructor's control. There are significant changes in the rankings when we adjust the SET score to eliminate the effects of variables beyond instructors' control. Nonetheless, when constructing the 95% percentage confidence interval of the predicted SET scores, we find that these changes are not statistically significant.

This paper has five sections including this introduction. In the next section, we present the data and the methodology employed in the analysis. The results are presented and discussed in section 3. In section 4, we analyze how the instructors' ranking changes when we

adjust the SET scores in order to eliminate either the possibility of instructor's manipulation through grade inflation or the effects of variables beyond the instructor's control. The last section concludes.

2. Methodology and Data

We obtained the data from Insper's (Institute of Education and Research) Academic Records Office. The data covered six semesters for the period from the second semester of 2005 to the first semester of 2008. It comprises 496 undergraduate courses offered during this period, taught by 101 different instructors. 63 observations were excluded from the sample, or 12,7% of the total, for three reasons: (i) the instructor taught only one time at the institution (32 observations in the original sample), (ii) the fraction of students enrolled in the class that answer the SET form was equal to 0% (3 observations) or greater than 100% (14 observations) and (iii) the number of students enrolled in class was smaller than 12 (25 observations).

It is important to point out that there are instructors who taught during the period analyzed more than one course in the same semester. This fact precludes the use of panel data techniques when using the instructor as the unity of analysis. In order to circumvent this problem, we consider the pair instructor/course as the unit of analysis. When the unity instructor/course occurred more than once in the same semester, the information related to this unity of observation was averaged. When averaged, the number of observations is equal to 363 and there are 130 pairs instructor/course.

There are two important characteristics of the data. The first is that Insper's students in each field (business or economics) must take the same courses in the first three of the four years of courses necessary to obtain the degree. Hence, in the first three years, they cannot choose either the instructor or the class, that is, they have to take the options offered. In the last year, students can select different course/instructor from the pool of offered elective disciplines. We conducted an F test to check if it is appropriate to pool together mandatory and elective courses.[4] The F-statistic is 25.71 (19 degrees of freedom) which is not significant at the usual level of significan-

[4] Following McPheerson *et al*. (2007), we tested the equality of parameters in mandatory and elective disciplines using specification in column 1 in Table 2 (see the appendix).

ce. It indicates that it is valid to pool the two groups of courses. Therefore, we conduct the empirical analysis combining data from both types of courses. The second characteristic is that, in the first three semesters, the courses are the same for students in the business or economics field and they are offered as joint courses. In this case, students cannot choose the instructor/class but are allocated by the institution, which mix economic and business students.

Insper hires individuals other than the instructors to distribute SET forms without announcement beforehand two times during the semester. They occur right before the mid-term and final exams. In our analysis, we use only the results obtained in the last evaluation. In this research, we use two dependent variables. The first one is the average of all answers in the SET form (hereafter referred as EVAL1), which is the variable used by the institution to evaluate the quality of instructor's teaching for purposes of promotion. EVAL1 ranges from 1 to 4, where a higher value indicates a better evaluation. The average score for EVAL1 for all courses was 3.32 and the minimum and maximum value were, respectively, 1.9 and 3.9. In Table 1 of the appendix, we present the descriptive statistics.

The second dependent variable (EVAL2) is calculated based on the answer to the following question, which is not used in the computation of EVAL1: "Considering the overall course and the instructor's performance, would you recommend this course with this instructor to a colleague?" The possible answers are no (value 1) or yes (value 2). EVAL2 is the average response and it obviously ranges from 1 to 2, where the closer is to 2 the better is the evaluation. The average value for EVAL2 for all courses was 1.84 and the minimum and maximum value were, respectively, 1.08 and 2.

Following the literature, we consider three groups of variables that can affect the SET score. They are related to the characteristics of the students, courses and instructors.

With respect to instructor's characteristics, we use several explanatory variables. The first two are related to instructor training programs. One is a dummy variable (CPCL) equal to 1 if the instructor had taken part in the Colloquium on Participant-Centered Learning at the Harvard Business School[5] and 0 otherwise. This training pro-

[5] For more details on this program, see http://www.exed.hbs.edu/programs/gcpcl/.

gram aims to help instructors to improve their effectiveness by learning from their teaching. The expected sign of this variable is positive as instructors learn new techniques and ways to improve their teaching. 17.6% of all instructors at Insper had taken part in this program. Another dummy variable is (PAAP), which is equal to 1 if the instructor had taken part in the PAAP program[6] and 0 otherwise. The PAAP program is one in which an instructor attends another instructor's class with the objective to identify problems, provide recommendations and suggestions in order to improve the teaching quality. Therefore, the expected sign of this variable is positive. 20.8% of all instructors had participated in this program.

We also control for the instructor's schooling. A dummy variable (PHD) is equal to 1 if the instructor has a PhD degree and 0 otherwise. The expected sign of this variable is positive as the instructor's knowledge and human capital increase with education. 76.2% of all instructors at Insper have a PhD degree. The others either have a master degree or a professional degree such as an MBA.

Another control variable is a dummy (GENDER) equal to 1 to male instructors and 0 otherwise. 82.2% of all instructors are male. It may exists some gender bias in the evaluation process, for example, due to discrimination or a different perception by the students of male *vis-à-vis* female instructor,[7] which makes unclear the sign of its coefficient. Another dummy variable is (FULL), which is equal to 1 or 0 if the instructor is, respectively, a full-time or part-time professor. The fraction of full-time professors is 27%. The sign of its coefficient is uncertain. Both types of instructors have other responsibilities rather than teaching.

One additional explanatory variable is the number of semesters teaching at Insper (EXP). The average number is equal to 3.6. This variable is a proxy for teaching experience, as we do not have the information of how long the instructor teaches at other institutions. The expected sign of its coefficient is positive as more experience in the classroom contributes to an increase in the teaching quality. In particular, as this variable counts only the number of semesters

6 PAAP stands for "professor attending another professor's class" in Portuguese.

7 Hamermesh and Parker (2005) indicate that beauty perception affects the SET score and its effect differs by instructor gender.

teaching at Insper, it may capture the instructor's adaptability to the institution's environment and student body. We also check if this learning gain reduces over time by the introduction of the EXP squared. The last variable is the instructor's age (AGE). The average age is 39.6 years old. Controlling for experience, the expected sign of its coefficient is negative due to different reasons: human capital depreciation, students' bias in favor of younger instructors and involvement in other activities rather than teaching such as administrative duties in the case of full-time professors.[8] We also check if the AGE squared is significant.

With respect to students' characteristics in each class, we use three explanatory variables. The first one is the actual average grade (GRADE). It ranges from 0 to 10 and its average score is 6.44, with minimum 2 and maximum 8.7. This variable may test the possibility that instructors can "buy" a better evaluation by giving higher grades.[9] Under this possibility, its coefficient is expected to be positive. As we do not have the average expected grade, which is more frequently used in the literature, this is the alternative employed.[10,11] The second variable is the fraction of students enrolled in class that answer the SET form (PRESP). The average value of PRESP is 60.8%. The expected sign of its coefficient is not clear. A high percentage of response may lead to lower SET scores either because the students are poorly satisfied with the instructors' performance and want to show their lack of appreciation or because a high fraction of low performing students answer the evaluation. The reverse may occur if a high percentage of response is indicative of student interest. The third variable is the fraction of female students in class (PFEM). The average value of PFEM is 27.9%. Again, the sign of its coefficient is unclear. The gender composition may affect the SET scores if male and female students have different standards when evaluating their instructors.

8 See discussion in McPherson *et al.* (2007).

9 This effect is of particular interest in the literature. See survey about this topic in McPherson *et al.* (2007).

10 Isely and Sing (2005) consider the relevant variable the difference between expected grade and the grades that students are used to receive. McPherson *et al.* (2007) argue that it is Amore appropriate to use the expected grade.

11 In the literature, there is some indication that the grade variable may be endogenous. See for example Seiver (1983) and Nelson and Lynch (1984). We return to this point in the next section.

With respect to the courses' characteristics, we use several explanatory variables. The first one is the number of students enrolled in class at the beginning of the semester (CSIZE). The sign of its coefficient is likely to be negative as the instructor provides less attention to any particular student the greater the class size and should be "penalized" by the students in the SET evaluation. CSIZE ranges from 13 to 115, and its average value is 56.5. Then, we use a dummy variable (MAND) equal to 1 if the course is mandatory and 0 otherwise. The percentage of mandatory courses in the sample is 84.5%. One should expect instructors teaching elective courses to be better evaluated by the students as the latter had the option to choose the course/instructor. Finally, we use a set of dummy variables to indicate whether the course belongs to the business degree (BUS) (21.9% of the total), to the economics degree (ECON) (17.8% of the total) or is a joint one (JOINT) (60.3%). The sign of their coefficients are unclear. The composition of the student body in class may affect the SET scores if economics and business students have different standards when evaluating their instructors.

We have a panel data and we consider two types of models, either a fixed-effect or a random-effects one. Both models control for the unobservable characteristics of the pairs instructor/course and time (from the 2nd semester of 2005 until 1st the semester of 2008). We test which model is the most appropriate. Hausman test for the sample indicates that the unobserved instructors' heterogeneity can be assumed to be uncorrelated with the explanatory variables, mentioned above, included in the analysis. The chi-square statistic is 8.84 (19 degrees of freedom), which is insignificant at any conventional level.[12]

Therefore, we concentrate the analysis in the results when the random-effects model is used, which is characterized by the following formulation:

$$Y_{it} = \alpha + u_i + \gamma_t + X_{it}\beta + \varepsilon_{it},$$

where: Y_{it} is the dependent variable (EVAL1 or EVAL2) of the pair instructor/course 'i' in semester 't'; α is a constant; u_i is the pair instructor/course specific effect; γ_t is the semester-specific effect; X_{it} is a vector that includes all the explanatory varia-

[12] We have conducted the test using specification in column 1in Table 2 (see the appendix).

bles mentioned above; β is a vector with the coefficients of interest; and ε_{it} is the error term and it is assumed to be well-behaved.

3. Results

In this section, we present the evidence of which factors affect the SET score. Table 2 in the appendix reports the estimation results using different specifications of the model. In column 1, we use the variable EVAL1 as the dependent variable and all the explanatory variables mentioned in the previous section, without the quadratic terms.

Before proceeding with the results, it is important to note that the variable EVAL1 is not truncated neither from below nor from above. It ranges from 1 to 4 and, as one can see in Table 1, in our sample, its minimum value is 1.9 and its maximum value is 3.9. Therefore, it is appropriate to use the FGLS econometric technique.

The results in column 1 indicate the following. With respect to the students' characteristics, there is only one that affects EVAL1: the GRADE. Hence, there is evidence that an instructor may be able to "buy" a better evaluation by inflating the students' grade. The coefficient of the variable GRADE is positive and significantly different from zero, but small. One point increase in GRADE in the 0-10 scale leads to an increase in the SET scores of 0.09 point. To give a better idea of this impact, consider two identical average classes with the exception of their average grades: one has the average grade of all classes[13] (6.47) and another has a grade one standard deviation lower (5.48). The instructor's SET score in latter would be 2.7% smaller than the former.[14]

There is no indication that male and female students have different standards when evaluating instructors, as the coefficient of the variable PFEM is not statistically different from zero. Finally, the fraction of students enrolled in the class that answer the SET form (PRESP) also is not relevant to explain the dependent variable EVAL1.

[13] The values in these comparisons are related to the adjusted sample for the pair instructor/course.

[14] As in McPherson *et al.* (2007), we employ a Hausman test (available from the authors), and we do not find evidence of endogeneity on the grade variable.

With respect to the courses' characteristics, the results are the following. The greater the number of students in class, the lower is the SET grade. In other words, the sign of the CSIZE's coefficient is negative and significantly different from zero. An additional student in class reduces the SET score in 0.003 point. To understand this effect better, consider two identical average classes with the exception of their sizes: one has 54.7 students in class (average number for all classes) and another has one standard deviation higher (75.64 students). The instructor's SET score in the latter would be 2% bigger than the former.

One can argue that it may have an endogeneity problem related to the class size, for example, in the case that students choose the courses taught by the best instructors. In order to check this possibility, we restrict our sample to the mandatory courses, the ones taught in the first three years of the four years courses of economics and business. The reason is that students cannot choose instructors or courses in the first three years of their courses. They are assigned by the institution. When we re-do the analysis using this restricted sample, the results do not change.[15]

There is no difference in terms of evaluation by students if the course is mandatory or elective. The coefficient of the dummy variable MAND is not statistically different from zero. This is a surprising result as one should expect students to evaluate better an instructor that he can choose. Finally, the coefficient of the dummy variable JOINT is positive and significantly different from zero. It indicates that the composition of the student body affects the SET score. In particular, classes with economics and business students evaluate better the instructors relatively to classes with only economics students. With respect to the instructors' characteristics, some variables are not statistically significant in the regression in column 1. The first one is the GENDER variable. There is no indication of discrimination or difference in perception with respect to the instructor's gender by the students. The second one is the status as full-time professor or the dummy variable FULL. This result is not surprising given that both types of instructors have other duties rather than teaching that should interfere in the same way their time allocation to class preparation. The coefficients of the variables AGE and PhD

[15] These results are available under request. We thank one of the referees for pointing out this possible endogenous problem.

are not statistically significant, as their expected signs were, respectively, negative and positive, as discussed in the previous section.

The coefficients of the two variables related to the training programs (PAAP and CPCL) are also not significantly different from zero. That is, there is no indication that students evaluate better the instructors who have passed by these two types of training programs. These are surprising results, as indicated by the discussion in the previous section. It is important to point out that, as the link between quality of teaching and SET scores is not clear, these results do not necessarily indicate that these programs are not capable of improving the quality of the instructors' class.

Depending on the way that instructors are selected to participate in both training programs (PAAP or CPCL), it could create an endogeneity problem. For example, if instructors with relatively lower SET scores are the ones selected to take part in the programs, then the dependent variable affects the dummy variables related to the variables PAAP and CPCL. We did not find a reasonable way to correct this potential problem. The difficulties are the following. One the one hand, there is no written policy that indicates how the instructors are actually selected into the programs. One the other hand, it is hard to find instrumental variables in this case, an exogenous variable with respect to the dependent variable and correlated with the variables PAAP and CPCL. In order to test the robustness of the other results in the regression, we estimated the models without the variables PAAP and CPCL and found that the other coefficients and its significance do not change.[16]

One variable related to the instructors' characteristics affects the SET scores. As expected, the longer the instructors' experience teaching at Insper, the greater is their evaluations. In other words, the coefficient of the variable EXP is positive and significantly different from zero. One additional semester of experience leads to an increase in the evaluation by 0.03 point. An example illustrates this effect. Consider two classes identical in all aspects but the instructors' experience. In the first one, the instructor has 3.4 semesters of experience, the average of all classes. In the second one, the instructor has one standard deviation lower than the average (0.9

[16] These results are available under request. We thank one of the referees for pointing out this possible endogenous problem.

semester). The instructor's SET score in former and the latter would be, respectively, equal to 3.3 and 3.2 points.

In column 2 in Table 2, we present a different specification of the random-effects model. It differs from the first column by the fact that there is one additional explanatory variable. It is the quadratic term of the variable EXP, which is negative and significantly different from zero. The coefficient of the variable EXP remains positive and significant. In other words, the first and second derivatives of the EVAL1 with respect to EXP are, respectively, positive and negative. Combining these two results, the empirical evidence suggests that instructors with more experience are better evaluated by the students but these gains reduce over time. There are no qualitative changes in the results under this new specification.

We tried some different specifications to the model in column 2. First, we introduced the quadratic terms of the variables CSIZE and AGE, but their coefficients were not significantly different from zero. Second, we included an interaction term between GENDER and PFEM to check if classes with a higher fraction of female students evaluate female instructors differently. We found no evidence of this effect. Third, as AGE and EXP are somewhat correlated, we run regressions with each variable separately but the results do not change. Finally, we extracted from the regression the variables EXP and its quadratic term and introduced the term AGE and its quadratic term. The idea behind this last change is that "AGE variables" may better reflect the professional experience than the "EXP variables". We do not find that the coefficients of the "AGE variables" to be significantly different from zero.[17]

Specifications in columns 3 and 4 in Table 2 are the same as in, respectively, columns 1 and 2. The only difference is that we use instead the dependent variable EVAL2.[18] Quantitatively speaking, the new results are not very different with some exceptions. The coefficient of the dummy variable JOINT is not statistically different from zero. Moreover, the coefficients of the variables GRADE and CSIZE are roughly, respectively, two and three times greater when the dependent variable is EVAL1. In fact, the coefficient of the

[17] These results of these different specifications are available from the authors under request.

[18] The variable EVAL2 is truncated from the right, as there are some instructors who receive the maximum possible score 2. However, the results do not change when we eliminate these observations from the sample.

variable CSIZE is not statistically different from zero in the model in column 3.

4. Rankings

The analysis in the previous section indicates that several factors affect the SET scores. Among these factors are variables under the control of the instructor, such as GRADE, or not, such as EXP and CSIZE. As a consequence, an instructor can receive a better evaluation either by manipulating his score through grade inflation or by the effects of variables that are beyond his control. In both cases, comparison of instructors without controlling for these possibilities may not be fair. In order to take into consideration these possibilities and adjust the SET scores accordingly, we construct three different rankings.[19] They are reported in Tables 3 and 4.

The benchmark case (ranking 1) is in Table 3. To obtain this first ranking, we do the following. We obtain the predicted SET score for each instructor for every time that he teaches a course by calculating the regression fitted value, using the estimated coefficients in column 2 in Table 2 and given the explanatory variables for each instructor. Note that this predicted value is not influenced by the instructor-specific random-effects. The results reported are the average fitted values over all semesters.

The second ranking is reported in Table 4. In this one, the procedure is the same as the one used to produce the benchmark case, with one exception. We replace the actual value of the explanatory variable GRADE of each instructor every time he teaches a course by the mean GRADE of the sample. Again, the reported results are the average fitted values over all semesters. By adjusting the ranking in this way, we eliminate the effects on the SET score of possible manipulation by the instructor through the grade inflation.

Comparison between rankings 1 and 2 indicates that they are very similar. Despite of the possibility of being able to "buy" higher scores by inflating students' grade, instructors in general are not able to change dramatically their positions in the ranking. The most significant change occurred with instructor I59 that moved from po-

[19] For similar adjustments in the literature, see Mason *et al.* (1995) and McPherson (2006).

sition 59 in the benchmark ranking to position 51 in ranking 2. In addition, instructor (I26) and instructors (I14, I36 and I51) moved down, respectively, 6 and 5 positions and instructor I21 moved up five positions.

The third ranking is reported in Table 4. Again, they are constructed in the same way as the benchmark case but with two exceptions. For each instructor and every time he teaches a course, we replace the actual values of two explanatory variables in the regression, CSIZE and EXP, by their respectively average values in the sample. The new ranking is formed by the average fitted values over all semesters. With these two adjustments, we basically eliminate the effects of variables beyond the instructor's control.

Rankings 1 and 3 have their similarities. For example, nine out of the ten top instructors in ranking 1 are also in the top ten positions in ranking 2. However, the differences between rankings 1 and 3 are greater than the ones between rankings 1 and 2. One indicator illustrates this fact. The sum of the absolute changes in positions of all instructors when one compares rankings 1 and 2 is equal to 104. This statistic is equal to 284 when the comparison is made between rankings 1 and 3. In fact, there are some dramatic changes in some instructors' positions in ranking 3 vis-à-vis ranking 1. For example, instructor I34 moves down 21 positions when one eliminates the effects of the variables CSIZE and EXP, which are not in his control. In contrast, instructors I31 and I40 move up, respectively, 14 and 13 positions.

We apply a more robust test to compare the rankings. We use the Wilcoxon Two Sample Test,[20] which is a nonparametric approach to check the null hypothesis of equality in the distributions behind two rankings. When comparing ranking 1 (the benchmark case) with ranking 2, one cannot reject the null hypothesis. The test p-value is equal to 0.2105. However, when comparing rankings 1 and 3, the null hypothesis is rejected at 5%, with the test p-value equal to 0.0121. These results are in line with the ones obtained above by simply comparing the positions of the instructors between rankings: rankings 1 and 2 are similar, whereas rankings 1 and 3 are not.

[20] For more details, see Wilcoxon (1945).

Tables 3 and 4 also show the predicted SET scores for all instructors in all three rankings with their respective 95% confidence interval. Despite the changes in positions in ranking 2 and in particular in ranking 3 with respect to ranking 1, the SET scores in all three rankings for all instructors are not statistically different.[21]

The last point we want to address is the following. Suppose that an institution establishes a threshold SET score, say 3.4,[22] such that instructors with scores greater or equal than this number are considered as having performed an outstanding job with important influences in their promotion status. A comparison of the predicted scores in rankings 1 and 2 suggests the following. No instructor who receives a score below 3.4 in ranking 1 (instructors I24 to I69) would pass this threshold in ranking 2, when the possibility of manipulating the score through grade inflation is taken into account and eliminated. At the same time, only instructor I23, who receives a score above 3.4 in ranking 1 (instructors I1 to I23), would have his score reduced to a level below to this threshold in ranking 2.

However, there are instructors with predicted scores below 3.4 (from instructor I23 with 3.39 to instructor I44 with 3.24 in ranking 2 in Table 4) whose value is not statistically different from 3.4, using the 95% confidence interval. Therefore, the use of the threshold 3.4 as the basis for promotion, without considering adjustments, should be used with cautious.

5. Conclusions

We estimated the factors that affect the SET scores, using a large panel data and a random-effects model in which it was possible to control for unobserved characteristics of the instructors as well as time-invariant ones. The results indicate that some variables influence the evaluation. They also seem robust to different specifications of the model.

There is evidence that an instructor may be able to 'buy' a better evaluation by inflating students' grade, though the effect is not

[21] We also calculated two different rankings, respectively, controlling for the variables CSIZE and EXP separately. They are also not statistically different from ranking 1.

[22] This value is the threshold value at Insper.

strong. As expected, the greater the number of students in class, the lower is the SET score. Moreover, instructors with more semesters of experience teaching at the institution are better evaluated by the students but these gains reduce over time. In addition, it is somehow surprising that the instructor's age and schooling do not affect the way students evaluate him.

Another result is the evidence that instructors who participated in training programs, designed to improve the quality of teaching, do not receive higher SET scores. One cannot easily conclude from these results that these programs are not capable of improving the quality of the instructors' class. The reason is that the link between quality and effectiveness of teaching and SET scores are not clear.

We construct different instructors' rankings by adjusting the SET scores, in order to eliminate either the possible manipulation by the instructor through grade inflation or the effects of variables beyond the instructor's control. There are significant changes in the rankings when we adjust the SET score to eliminate the effects of variables beyond instructors' control. Nonetheless, when constructing the 95% percentage confidence interval of the predicted SET scores, we find that these changes are not statistically significant. One important policy implication is that the use of a given threshold SET score as the basis for promotion, without considering adjustments, should be used with cautious.

Finally, our qualitative results are very similar to the ones obtained in the closest papers to ours, McPherson (2006) and McPherson *et al*. (2007): instructors seem to be able to 'buy' better evaluation scores, class size and experience seem to affect the instructors SET scores and it is important to adjust the instructors' ranking. Hence, different social and cultural environments (higher education institutions in Brazil and the US) did not seem to affect the instructors' and students' behaviors.

Appendix

Table 1 - Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
SET Variables					
EVAL1	433	3.3	0.3	1.9	3.9
EVAL2	433	1.8	0.2	1.1	2.0
Student's Characteristics					
GRADE	433	6.4	1.0	2.0	8.7
PRESP	433	60.8%	16.4%	8.0%	93.6%
PFEM	433	27.9%	7.9%	6.8%	52.5%
Courses' Characteristics					
CSIZE	433	56.5	20.7	13.0	115.0
MAND	433	84.5%	36.2%	0.0%	100.0%
BUS	433	21.9%	41.4%	0.0%	100.0%
ECON	433	17.8%	38.3%	0.0%	100.0%
JOINT	433	60.3%	49.0%	0.0%	100.0%
Instructors' Characteristics					
GENDER	433	82.2%	38.3%	0.0%	100.0%
FULL	433	27.0%	44.5%	0.0%	100.0%
EXP	433	3.6	2.4	0.0	14.4
AGE	433	39.6	8.0	24.9	64.3
PHD	433	76.2%	42.6%	0	1
PAAP	433	20.8%	40.6%	0	1
CPCL	433	17.6%	38.1%	0	1

Table 2 - Random-Effects' FGLS Estimates [1]

	Dependent Variable			
Explanatory variables	EVAL1	EVAL1	EVAL2	EVAL2
Student's Characteristics				
GRADE	0.091***	0.078***	0.049***	0.039***
	(0.000)	(0.002)	(0.000)	(0.004)
PRESP	0.000	0.000	0.000	-0.000
	(0.752)	(0.919)	(0.922)	(0.846)
PFEM	-0.001	-0.001	-0.001	-0.001
	(0.627)	(0.610)	(0.343)	(0.331)
Courses' Characteristics				
CSIZE	-0.003***	-0.003***	-0.001	-0.001*
	(0.003)	(0.001)	(0.191)	(0.096)
MAND	0.058	0.056	0.001	0.000
	(0.433)	(0.449)	(0.982)	(0.993)
BUS	0.052	0.060	0.032	0.037
	(0.580)	(0.522)	(0.518)	(0.443)
JOINT	0.150*	0.158**	0.046	0.052
	(0.066)	(0.050)	(0.251)	(0.179)

(Continued)

Instructors' Characteristics				
GENDER	0.066	0.074	0.036	0.042
	(0.372)	(0.306)	(0.388)	(0.302)
FULL	-0.032	0.006	-0.032	-0.005
	(0.650)	(0.929)	(0.375)	(0.892)
EXP	0.028**	0.080***	0.019**	0.056***
	(0.036)	(0.002)	(0.021)	(0.000)
EXP2		-0.005**		-0.004***
		(0.011)		(0.004)
AGE	-0.000	0.002	-0.001	-0.000
	(0.972)	(0.713)	(0516)	(0.947)
PHD	0.006	-0.030	0.026	-0.001
	(0.931)	(0.662)	(0.527)	(0.988)
PAAP	-0.036	-0.006	-0.060	-0.037
	(0.587)	(0.934)	(0.107)	(0.330)
CPCL	-0.010	-0.012	-0.040	-0.041
	(0.868)	(0.849)	(0.238)	(0.228)
Number of obs.	363	363	363	363
Wald chi2 (d.f.)	87.32	92.12	42.41	53.89
	(0.000)	(0.000)	(0.000)	(0.000)
R2 overall	0.159	0.206	0.1302	0.202

[1] The equations include time dummies for each semester between 2nd semester of 2005 until 1st semester of 2008. P-values based on White robust standard-errors in parentheses.
*, **, *** denote rejection of the null hypothesis at the 10%, 5% and 1% level respectively.

Table 3 - Instructors' Ranking: Estimates and 95% Confidence Intervals

Ranking 1 - Benchmark

Instructor	Lower Bound	Point Estimate	Upper Bound
1	3.52	3.65	3.79
2	3.45	3.64	3.83
3	3.51	3.64	3.77
4	3.46	3.63	3.81
5	3.42	3.61	3.80
6	3.46	3.59	3.72
7	3.36	3.53	3.69
8	3.32	3.52	3.71
9	3.37	3.50	3.64
10	3.23	3.49	3.74
11	3.33	3.48	3.63
12	3.31	3.48	3.66
13	3.32	3.45	3.59
14	3.26	3.44	3.63
15	3.27	3.44	3.61
16	3.32	3.43	3.54
17	3.25	3.43	3.61
18	3.31	3.42	3.54
19	3.27	3.42	3.57
20	3.24	3.41	3.58
21	3.29	3.41	3.53
22	3.21	3.40	3.59
23	3.28	3.40	3.52
24	3.23	3.39	3.55
25	3.23	3.38	3.53
26	3.21	3.37	3.54
27	3.22	3.37	3.51
28	3.21	3.37	3.52
29	3.22	3.37	3.51
30	3.17	3.36	3.55
31	3.17	3.35	3.53
32	3.17	3.33	3.49
33	3.20	3.33	3.46
34	3.16	3.33	3.49
35	3.18	3.32	3.47
36	3.15	3.32	3.49
37	3.16	3.32	3.47
38	3.18	3.31	3.44
39	3.18	3.31	3.44
40	3.17	3.31	3.45
41	3.13	3.30	3.48
42	3.18	3.30	3.42
43	3.17	3.29	3.41
44	3.05	3.25	3.45

(Continued)

Ranking 1 - Benchmark			
Instructor	**Lower Bound**	**Point Estimate**	**Upper Bound**
45	3.10	3.24	3.37
46	3.06	3.22	3.39
47	3.05	3.21	3.36
48	3.07	3.20	3.34
49	3.03	3.20	3.36
50	3.04	3.19	3.34
51	2.99	3.19	3.38
52	3.03	3.18	3.33
53	3.08	3.18	3.28
54	3.01	3.16	3.32
55	3.00	3.16	3.32
56	3.00	3.15	3.31
57	2.97	3.13	3.29
58	2.92	3.12	3.33
59	2.89	3.12	3.35
60	2.93	3.10	3.27
61	2.82	3.08	3.35
62	2.82	3.02	3.22
63	2.82	3.02	3.22
64	2.83	3.01	3.19
65	2.85	3.00	3.15
66	2.81	2.96	3.11
67	2.67	2.87	3.07
68	2.67	2.85	3.03
69	2.43	2.61	2.78

Table 4 - Adjusted Rankings: Estimates and 95% Confidence Intervals

Ranking 2: Control by GRADE				Ranking 3: control by CSIZE and EXP			
Instructor	**Lower Bound**	**Point Estimate**	**Upper Bound**	**Instructor**	**Lower Bound**	**Point Estimate**	**Upper Bound**
3	3.52	3.64	3.77	1	3.54	3.68	3.82
2	3.45	3.64	3.83	3	3.53	3.66	3.79
1	3.49	3.63	3.76	2	3.42	3.61	3.80
4	3.45	3.62	3.79	4	3.42	3.59	3.76
5	3.42	3.61	3.79	5	3.40	3.59	3.78
6	3.46	3.58	3.71	6	3.44	3.57	3.71
7	3.35	3.52	3.68	7	3.42	3.57	3.72
9	3.38	3.52	3.65	8	3.37	3.56	3.74
12	3.35	3.51	3.67	11	3.41	3.55	3.69
8	3.30	3.49	3.68	9	3.35	3.49	3.63
11	3.32	3.47	3.62	14	3.27	3.46	3.65

(Continued)

Ranking 2: Control by GRADE				Ranking 3: control by CSIZE and EXP			
Instructor	Lower Bound	Point Estimate	Upper Bound	Instructor	Lower Bound	Point Estimate	Upper Bound
13	3.31	3.45	3.59	19	3.29	3.45	3.62
10	3.20	3.45	3.69	12	3.28	3.45	3.63
15	3.27	3.44	3.61	13	3.31	3.44	3.58
16	3.32	3.43	3.54	10	3.19	3.44	3.69
21	3.32	3.42	3.53	18	3.32	3.44	3.56
18	3.31	3.42	3.54	31	3.27	3.43	3.60
20	3.26	3.42	3.59	26	3.26	3.43	3.61
14	3.23	3.41	3.60	29	3.28	3.43	3.58
17	3.23	3.41	3.58	30	3.22	3.42	3.62
22	3.21	3.40	3.59	25	3.26	3.42	3.58
19	3.25	3.40	3.55	16	3.30	3.42	3.54
23	3.28	3.39	3.51	15	3.24	3.42	3.60
25	3.24	3.38	3.53	36	3.26	3.40	3.55
29	3.24	3.38	3.52	21	3.28	3.40	3.53
24	3.23	3.38	3.53	17	3.21	3.39	3.57
31	3.19	3.37	3.54	40	3.24	3.38	3.52
28	3.20	3.36	3.52	20	3.18	3.38	3.58
27	3.21	3.36	3.50	22	3.18	3.38	3.58
34	3.18	3.35	3.53	23	3.26	3.38	3.50
30	3.16	3.35	3.54	24	3.21	3.36	3.52
26	3.19	3.35	3.50	27	3.20	3.35	3.50
33	3.21	3.34	3.46	35	3.20	3.34	3.49
35	3.20	3.33	3.47	28	3.17	3.34	3.52
32	3.17	3.33	3.49	37	3.17	3.33	3.50
37	3.18	3.33	3.48	32	3.15	3.32	3.50
39	3.20	3.32	3.44	33	3.18	3.32	3.45
38	3.19	3.32	3.45	41	3.13	3.31	3.50
40	3.16	3.30	3.44	42	3.18	3.31	3.44
43	3.18	3.29	3.41	38	3.18	3.31	3.44
36	3.11	3.29	3.47	45	3.17	3.31	3.44
42	3.18	3.29	3.41	39	3.17	3.30	3.44
41	3.12	3.28	3.45	43	3.16	3.28	3.40
44	3.03	3.24	3.44	46	3.09	3.27	3.44
45	3.09	3.23	3.36	51	3.11	3.26	3.42
46	3.06	3.22	3.39	44	3.07	3.26	3.45
48	3.08	3.21	3.34	50	3.11	3.26	3.40
49	3.04	3.21	3.37	54	3.11	3.25	3.39
47	3.04	3.20	3.36	55	3.06	3.23	3.40
50	3.05	3.20	3.34	48	3.08	3.22	3.36
59	2.91	3.19	3.46	52	3.05	3.19	3.34
52	3.04	3.18	3.33	57	3.02	3.19	3.36
53	3.08	3.18	3.29	49	3.02	3.19	3.36
55	3.02	3.18	3.34	47	3.02	3.19	3.36

(Continued)

Ranking 2: Control by GRADE				Ranking 3: control by CSIZE and EXP			
Instructor	Lower Bound	Point Estimate	Upper Bound	Instructor	Lower Bound	Point Estimate	Upper Bound
54	3.01	3.17	3.32	34	3.02	3.18	3.33
51	2.96	3.16	3.36	53	3.06	3.17	3.28
56	2.98	3.14	3.29	56	3.00	3.17	3.34
58	2.92	3.12	3.32	58	2.98	3.16	3.34
57	2.95	3.11	3.27	61	2.93	3.15	3.37
60	2.91	3.08	3.25	60	2.91	3.09	3.26
61	2.79	3.06	3.32	65	2.89	3.05	3.20
62	2.84	3.04	3.24	64	2.87	3.04	3.21
64	2.84	3.01	3.19	66	2.88	3.03	3.17
65	2.85	3.00	3.15	63	2.81	3.01	3.21
63	2.80	3.00	3.20	62	2.77	2.99	3.21
66	2.80	2.95	3.11	68	2.73	2.92	3.11
67	2.68	2.88	3.07	67	2.73	2.92	3.11
68	2.70	2.87	3.04	59	2.37	2.88	3.40
69	2.46	2.62	2.78	69	2.48	2.65	2.82

References

Becker, W., and Watts, M. (1999). "How Departments of Economics Evaluate Teaching". *American Economic Review*. Papers and Proceedings 89 (2): 344-49.

Gramlich, E. and Greenlee, G. (1993). "Measuring Teaching Performance". *The Journal of Economic Education* 24(1): 3-13.

Hamermesh, D. and Parker, A. (2005). "Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity". *Economics of Education Review* 24: 369-376.

Isely, P. and Singh, H. (2005). "Do Higher Grades Lead to Favorable Student Evaluations?" *The Journal of Economic Education* 36(1): 29-42.

Mason, P., Steagall, J. and Fabritius, M. (1995). "Student Evaluations of Faculty: A New Procedure for Using Aggregate Measures of Performance". *Economics of Education Review* 14 (4): 403-16.

McPherson, M. (2006). "Determinants of How Students Evaluate Teachers". *The Journal of Economic Education* 37(1): 3-20.

McPherson, M., Jewell, R. and Kim, M. (2007). "What Determines Student Evaluation Scores? A Random Effects Analysis of Undergraduate Economic Classes". *mimeo*.

Nelson, J. and Lynch, K. (1984). "Grade Inflation, Real Income, Simultaneity, and Teaching Evaluations". *The Journal of Economic Education* 15 (Winter): 21-37.

Seiver, D. (1983). "Evaluations and Grades: A Simultaneous Framework". *The Journal of Economic Education* 14 (Summer): 332-38.

Soper, J. (1973). "Soft Research on a Hard Subject: Student Evaluations Reconsidered". *The Journal of Economic Education* 5(1): 22-26.

Wilcoxon, F. (1945). "Individual comparisons by ranking methods". *Biometrics* 1: 80-83.