

Teacher professionalism and student evaluation of teaching: will better teachers receive higher ratings and will better students give higher ratings?

Pieter Spooren* and Dimitri Mortelmans

University of Antwerp, Belgium

The use of student evaluations of teaching performance has been an important but controversial tool in the improvement of teaching quality during the past few decades. Although student evaluations of teaching are implemented in many faculties, not everyone is convinced of the desirability and utility of these ratings. In this paper, we present the results of a study with regard to the existence of a higher-order factor that might influence students' perceptions of teaching and, thus, explain the variance in teacher rating scales. A second question concerns the effect of students' grades on teacher ratings and of other factors influencing this relationship (for instance, students' overall grades).

Keywords: Teacher evaluation; Evaluation methods; Students' evaluation of teaching; Validity studies

Introduction

Student evaluation of teacher performance has been an important but controversial tool in the improvement of teaching quality during the past few decades. Nevertheless, student ratings of instruction in higher education are not considered a recent phenomenon. Marsh (1987) and Wachtel (1998) report that the first 'teacher rating scale' was published in 1915, and that the first wave of studies of students' evaluations of teacher effectiveness were written in the 1920s (cf. Guthrie's and Remmer's pioneering work). The 'golden age of research on student evaluations', however, must be situated in the

*Corresponding author. Faculty of Political and Social Sciences, University of Antwerp, Universiteitsplein 1, B-2610 Wilrijk-Antwerpen, Belgium. Email: pieter.spooren@ua.ac.be
Phone: +32/3/820.28.33; Fax: +32/3/820.28.82

1970s, when much research examined the utility and validity of student ratings of instruction (see Centra, 1993).

Kulik (2001) states that the initial aim of student evaluations of teaching served two goals: mapping the quality of teaching in faculties/universities, and providing information and help to instructors in order to improve their teaching. Nowadays, student ratings are also set up and used in administrative decision-making, informing students concerning the selection of courses, curriculum development, external quality care and research on teaching (see e.g. Marsh, 1987; McKeachie, 1997). Although the implementation of student evaluations of teaching was applauded in many faculties, not everyone is convinced of the desirability and the utility of these ratings. Supporters argue that evaluative judgements on a regular basis have a strong positive influence on the improvement of instructional skills. It is logical that students, who indeed 'enjoy' the teaching and instruction, are involved in this form of quality care. Theall and Franklin (2001), for instance, state that:

Students spend a full term in the course, observe the instructor in class and in interactions with students, and can accurately judge what or how much they have learned with respect to their knowledge at entry. Students can report the frequencies of teacher behaviours, the amount of work required, and the difficulty of the material. They can answer questions about the clarity of lectures, the value of readings and assignments, the clarity of the instructor's explanations, the instructor's availability and helpfulness, and many other aspects of the teaching and learning process. No one else is as qualified to report on what transpired during the term simply because no one else is present for as much of the term. (p. 48)

However, opponents call into question student ratings of instruction since they have their doubts with regard to the validity of student perceptions of teaching (Sproule, 2002), and consider student ratings as 'meaningless quantification' and leading to 'personality contests' (see Kulik, 2001), instead of being valuable systems that measure teacher effectiveness. With regard to this, Marsh (1984) states that 'opinions about the role of students' evaluations vary from "reliable, valid and useful" to "unreliable, invalid and useless"' (p. 708). Kulik (2001) concurs, and notes that 'one might suppose that the research studies on ratings are similar to many other studies in education: conflicting, confusing and inconclusive' (p. 10). After all, the evaluation of teachers by their students has been widespread on college campuses for many years. Despite the intensive use of evaluation questionnaires (Chen & Hoshower, 2003, Wolfer & Johnson, 2003), assessing teaching is far from evident in many cases: 'Student ratings of university teachers have been common for at least thirty years, but it is a rare campus where they are accepted with equanimity' (Knapper, 2001, p. 3).

It is certain that most researchers believe that the results of student ratings provide evaluators with valid, reliable and valuable data concerning the quality and effectiveness of teaching (Penny, 2003). Marsh (1987) stated, as a result of an extensive review of the research literature, that student evaluation of teaching is probably the only indicator for teaching effectiveness of which validity has been proved this thoroughly. In addition, Centra (2003) refers to the large amount of research concerning this method of evaluation which, in general, shows that student evaluations are: (a) reliable and stable; (b) valid when compared with student

learning and other indicators of effective teaching; (c) only multidimensional in terms of what they assess; (d) useful in improving teaching; and (e) only minimally affected by various course, teacher or student characteristics that could bias results (see p. 496). Concerning this last finding, however, some authors have pointed to the existence of factors—concerning the teachers, the students and the course—influencing (or better, biasing) the results of student evaluations—e.g. gender, class size, grade expectations, teacher rank and experience, etc. (for an extensive overview see Marsh, 1987; Wachtel, 1998; Ckonko *et al.*, 2002). Supporters of student evaluations have therefore spoken (with or without reason) of a ‘witch hunt’ for potential bias in student ratings (Marsh, 1987; Theall & Franklin, 2001). In this study, we focus on a hot topic with regard to potential bias, namely the ‘grading leniency’ hypothesis. In what follows, we give a brief summary of the debates concerning this theme and discuss the results of our empirical evaluation research.

As a result of his overview of several research studies, Feldman (1976, 1997) reports the existence of a modest but significant correlation between (expected) grading and student ratings of teacher effectiveness of between 0.10 and 0.30, while Centra (2003) believes the correlation averages are close to 0.20. This seems only logical, but there exists a lasting discussion with regard to the interpretation of the association. Marsh (1987) suggests three possible explanations. The *grading leniency hypothesis* proposes that a teacher can ‘buy’ better student ratings by giving higher grades: ‘instructors who give higher-than-deserved grades will be rewarded with higher-than-deserved student ratings’ (p. 317). In this case, we could speak of serious bias, which exists when ‘a student, teacher or course characteristic affects the evaluations made, either positively or negatively, but is unrelated to any criteria of good teaching, such as increased student learning’ (Centra, 2003, p. 498). The *validity hypothesis* arises from the idea that better teachers make their students work harder and learn more, which leads to better learning results: good ratings reflect good learning. This implies that the correlation between (expected) grades and student ratings supports the validity of student evaluations of teaching (Marsh, 1987). The *student characteristic hypothesis* suggests the influence of pre-existing student characteristics on grades, teaching and student ratings, which makes the grade effect somewhat ‘false’.

Despite the fact that most studies supporting the grading leniency hypothesis (e.g. Greenwald & Gillmore, 1997a) were partly or fully refuted due to their methodological, statistical and/or theoretical weaknesses (Marsh, 1987; Marsh & Roche, 2000; Centra, 2003), there still exists evidence for this hypothesis (e.g. Brodie, 1998; Krautmann & Sander, 1999). Greenwald and Gillmore (1997b) reported that workload plays an important role too; and their research showed a negative correlation with (expected) grading. The conclusion, then, would appear to be that teachers who give high gradings and ask for little work from their students should not worry about student ratings.

However, experts and researchers on student evaluations of teaching agree on both the theory and empirical findings of the validity hypothesis whether or not combined with the student characteristic hypothesis—e.g. prior subject interest). Still, the threat of possible biases has not been suppressed. Contrary to what other researchers have suggested, we are convinced of the necessity of permanent research with regard to this

matter. The main reason for our tenacity concerns the use of student evaluations of teaching in both promotion and tenure decisions. As Penny (2003) argues, the use of student evaluations is seen as a key indicator in quality monitoring. In such a way, it is unlikely that student evaluations of teacher effectiveness will disappear soon, but, on the contrary, will be applied by administrators and policy-makers in the context of higher education. Thus, if student ratings are used in delicate issues like promotions and tenure decisions, we had better continue the quest both for valid instruments and conclusive evidence concerning possible biasing factors in student evaluations.

Objectives

In this paper, we present and discuss the results of research on student evaluation of teacher effectiveness. The initial goal of the study was to construct a valid evaluation questionnaire that allows students to share their experiences and appreciation concerning the courses they attended with their instructors. The shortcomings of the evaluation instruments we examined resulted in a research project aimed at developing a theory-based and thoroughly validated instrument that could be used in a wide variety of contexts (Mortelmans & Spooren, 2005). During this study, further questions concerning the correlation between grades and student ratings arose. Therefore, we decided to explore the data we gathered more profoundly in order to gain further insight into students' perspectives of good teaching and possible bias influencing student ratings. The main research questions thus concern the correlation between grading and student evaluation of teaching and tracing other factors that more or less seriously influence this relationship. First, we examined the existence of a higher-order factor that might influence student perceptions of teaching and, thus, explain much of the variance in the scales we placed in the evaluation instrument. A second question concerns the effect of students' grades in the course and their evaluation of the teaching in the course. In a third step, we tested the possible effect of some other factors—e.g. the students' overall grades—on student ratings of teaching skills.

Method

The evaluation instrument used in this study was constructed in a three-phased research project. A first step in measuring qualities of education is the determination of the minimal requirements of a sound educational practice. During this phase, we reviewed the literature to distinguish and define the various characteristics of teaching in relation to the assessment of the quality of education in university courses. In total, eight main dimensions were distinguished and further elaborated in 22 sub-dimensions. Based on the empirical translation of the theory, we formulated at least six Likert items for measurement of each sub-dimension. In this way, we built up an extensive test questionnaire with 165 items, representing all sub-dimensions. All the items were measured on a six-point scale (Strongly disagree to Strongly agree).

In the second phase, students from six different faculties ($N = 433$) were asked to rate a course they attended the previous academic year using the elaborated test

questionnaire. Using different statistical tools, we checked how well the items measured each of the 22 sub-dimensions. Items that did not comply with the minimal criteria of reliability and validity were deleted. After an initial exploration using Cronbach's alpha testing and factor analysis, the test questionnaire was cut back to 31 items that represented 10 of the 22 original sub-dimensions. Subsequently, we introduced these items in a confirmatory factor analysis using the Lisrel program. In our view, confirmatory factor analysis was best placed to extensively test the validity and reliability of the different scales. Table 1 summarizes the results of this analysis.

Table 1. Remaining scales in the confirmatory factor analysis

	Standardized factor loadings	<i>t</i> -test	Probability <i>t</i> -test	<i>R</i> ²	Variance extracted
F1: clarity of objectives ($\rho_c = 0.96$)					0.88
Item 11	0.92	55.32	0.001	0.84	
Item 12	0.98	79.12	0.001	0.97	
Item 13	0.92	52.40	0.001	0.84	
F2: value of subject-matter ($\rho_c = 0.90$)					0.75
Item 21	0.91	53.17	0.001	0.82	
Item 22	0.82	36.90	0.001	0.67	
Item 23	0.87	40.07	0.001	0.76	
F3: build-up of subject-matter($\rho_c = 0.91$)					0.77
Item 31	0.86	41.51	0.001	0.74	
Item 32	0.95	53.10	0.001	0.91	
Item 33	0.82	36.37	0.001	0.67	
F4: presentation skills ($\rho_c = 0.94$)					0.85
Item 41	0.94	75.66	0.001	0.89	
Item 42	0.91	55.03	0.001	0.83	
Item 43	0.92	62.60	0.001	0.84	
F5: harmony organization course—learning process ($\rho_c=0.85$)					0.58
Item 51	0.77	31.48	0.001	0.59	
Item 52	0.85	38.44	0.001	0.72	
Item 53	0.67	23.62	0.001	0.45	
F6: (course materials) contribution to understanding the subject-matter ($\rho_c =$ 0.91)					0.72
Item 61	0.90	45.48	0.001	0.81	
Item 62	0.84	38.42	0.001	0.70	
Item 63	0.90	47.31	0.001	0.81	
Item 64	0.75	31.60	0.001	0.56	
F7: course difficulty ($\rho_c = 0.86$)					0.68
Item 71	0.94	47.68	0.001	0.88	
Item 72	0.79	31.47	0.001	0.63	
Item 73	0.73	30.07	0.001	0.53	
F8: help of the teacher during the learning process ($\rho_c = 0.91$)					0.77
Item 81	0.86	44.37	0.001	0.73	
Item 82	0.89	42.38	0.001	0.79	
Item 83	0.88	44.87	0.001		

Table 1. (continued)

	Standardized factor loadings	<i>t</i> -test	Probability <i>t</i> -test	<i>R</i> ²	Variance extracted
F9: authenticity of the examination(s) ($\rho_c = 0.83$)					0.63
Item 91	0.68	22.79	0.001	0.47	
Item 92	0.84	33.79	0.001	0.71	
Item 93	0.84	34.53	0.001		
F10: formative examination(s) ($\rho_c = 0.83$)					0.61
Item 101	0.80	28.94	0.001	0.63	
Item 102	0.86	30.40	0.001	0.73	
Item 103	0.69	21.32	0.001	0.48	

Notes: Scale composite reliability: $\rho_c = \left[\left(\sum \lambda_i \right)^2 \text{var}(\xi) \right] / \left[\left(\sum \lambda_i \right)^2 \text{var}(\xi) + \sum \theta_{ii} \right]$ (Bagozzi & Yi, 1988,

p. 80). Fit statistics for confirmatory factor analysis of 31 indicators for 10 constructs: $\chi^2(701) = 1102.58, \rho = 0.00$; GFI = 0.96; CFI = 0.97; PNFI = 0.79; NNFI = 0.97; RMSEA = 0.039.

The confirmatory factor model, presented in Table 1, provides a good fit to the data (RMSEA = 0.039, NNFI = 0.97, CFI = 0.97). The χ^2 test of exact fit is significant, where the objective is to achieve a non-significant *p*-value. However, Hatcher (1994, p. 289) indicates that a significant χ^2 does not render the confirmatory factor analysis model inadequate. The χ^2 test indicates that the ratio of the value and the degree of freedom is lower than 2 (1.57), and that the test is within acceptable limits (Marsh *et al.*, 1988). In terms of the constructs, convergent validity is evidenced by the large and significant loadings of the items on their posited indicators. Further evidence of convergent validity is shown in Table 2. None of the correlations between the latent constructs was too high to challenge the convergent validity of the construct.

Discriminant validity is indicated because the confidence interval (\pm two standard errors) around the correlation estimate between any two latent constructs never

Table 2. Construct correlation matrix

Construct	F1	F2	F3	F4	F5	F6	F7	F8	F9
F2	0.68								
F3	0.79	0.81							
F4	0.75	0.65	0.73						
F5	0.68	0.59	0.67	0.65					
F6	0.73	0.66	0.84	0.72	0.57				
F7	0.79	0.62	0.68	0.61	0.56	0.65			
F8	0.83	0.63	0.71	0.68	0.78	0.67	0.80		
F9	0.43	0.38	0.42	0.48	0.68	0.29	0.23	0.36	
F10	0.40	0.18	0.40	0.29	0.37	0.36	0.31	0.41	0.08

scores 1.0 (Anderson & Gerbing, 1988, p. 416). The variance extracted test also shows the discriminant validity of our constructs; this test compares the variance extracted from two latent constructs with the square of the correlation between the two constructs (Fornell & Larcker, 1981). Discriminant validity is indicated when the explained variance is greater than the squared correlation. We compared all pairs of factors and they showed acceptable variance extracted.

Because phase 2 consisted of a test case in which 165 items were reduced to 31 items, we presented the selected items in a third phase as an independent instrument to a new sample of students. As in the second phase, the students were asked to evaluate a particular course they attended during the last academic year, but this time they filled in the questionnaire on two occasions (with an interval of one or two weeks); in order to link the questionnaire from both time points, they were asked to provide some extra identification details. The evaluation results for eight courses (completed by 566 students) allowed us to execute a number of reliability and stability tests, as shown in Table 3.

The results of these tests show that the 31 items included in our test instrument were reliable. The only deviation is the reoccurrence of rather low kappa values. This might be related to the use of the six-point scales on which the items were scored. If we reduce the six categories to three ('disagree', 'neutral' and 'agree'), we find that

Table 3. Validity and reliability tests of the evaluation instrument

	Internal consistency (Cronbach's alpha)		Cohen's kappa	Kendall's tau	Spearman's corr.
	<i>t</i> 1	<i>t</i> 2	(<i>t</i> 1 - <i>t</i> 2) (<i>p</i> = 0.00)	(<i>t</i> 1 - <i>t</i> 2) (<i>p</i> = 0.00)	(<i>t</i> 1 - <i>t</i> 2) (<i>p</i> = 0.00)
F1: clarity of objectives	0.781	0.782			
Item 11			0.337	0.601	0.682
Item 12			0.390	0.516	0.613
Item 13			0.366	0.556	0.652
F2: value of subject-matter	0.714	0.758			
Item 21			0.373	0.510	0.688
Item 22			0.404	0.454	0.510
Item 23			0.389	0.559	0.638
F3: build-up of subject-matter	0.668	0.729			
Item 31			0.334	0.283	0.487
Item 32			0.286	0.502	0.548
Item 33			0.397	0.573	0.635
F4: presentation skills	0.898	0.875			
Item 41			0.550	0.793	0.861
Item 42			0.567	0.756	0.814
Item 43			0.401	0.593	0.673

Table 3. (continued)

	Internal consistency (Cronbach's alpha)		Cohen's kappa	Kendall's tau	Spearman's corr.
	<i>t</i> 1	<i>t</i> 2	(<i>t</i> 1 – <i>t</i> 2) (<i>p</i> = 0.00)	(<i>t</i> 1 – <i>t</i> 2) (<i>p</i> = 0.00)	(<i>t</i> 1 – <i>t</i> 2) (<i>p</i> = 0.00)
F5: harmony organization course—learning process	0.663	0.760			
Item 51			0.438	0.549	0.608
Item 52			0.279	0.497	0.571
Item 53			0.251	0.514	0.606
F6: (course materials) contribution to understanding the subject-matter	0.875	0.856			
Item 61			0.548	0.738	0.797
Item 62			0.356	0.629	0.745
Item 63			0.453	0.602	0.713
Item 64			0.440	0.512	0.567
F7: course difficulty	0.835	0.858			
Item 71			0.473	0.698	0.761
Item 72			0.376	0.601	0.666
Item 73			0.435	0.692	0.756
F8: help of the teacher during the learning process	0.765	0.730			
Item 81			0.408	0.509	0.556
Item 82			0.228	0.558	0.636
Item 83			0.297	0.550	0.643
F9: authenticity of the examination(s)	0.858	0.833			
Item 91			0.351	0.543	0.609
Item 92			0.336	0.552	0.611
Item 93			0.281	0.497	0.581
F10: formative examination(s)	0.779	0.833			
Item 101			0.468	0.705	0.793
Item 102			0.403	0.528	0.589
Item 103			0.381	0.613	0.709

kappa values are remarkably higher. In other words, working with fewer scale values increased the reliability of the instrument. But on the other hand, working with more scale values increased the sensitivity of the instrument. Because the kappa values were only affected by the number of answer categories, we can conclude that retest reliability is not objected. All 10 latent dimensions remained stable in the third phase.

Sample

Because the students in the third phase were asked to give us further identification details, we obtained access to some interesting background variables, which could be used in the present study. These variables were: (1) final grade in the course (between 0 and 20), (2) total grade over all courses (%), (3) class size and (4) gender. These extra variables were added to the evaluation results of 222 students, which could be used for further investigation. We acknowledge that this is a rather small sample; our research must thus be seen as exploratory, although we found some interesting correlations that should be further examined in later research projects.

Results

Halo effect?

According to Pike (1999), there is growing evidence that student evaluations of teaching may be an artefact of a ‘constant error of the halo’ (p. 61). In other words, it is argued that one underlying factor more or less seriously influences students’ perceptions of teaching. This factor might mask relationships between educational outcomes and college experiences (p. 63); Shevlin *et al.* (2000), for example, found that their teacher charisma factor explained 69% and 37% of the variation in their ‘lecturer ability’ and ‘module attributes’ factors respectively, and they suggest that a central higher-order factor (‘charisma’) influences student ratings of teaching effectiveness. Since we designed 10 Likert scales that measured 10 different sub-dimensions of teaching practice (for which convergent and discriminant validity was proved), we decided to put it to the test in the first phase of our study in a confirmatory factor analysis, by assuming an underlying factor explaining the variation in our sub-dimensions. Table 4 shows

Table 4. The teacher professionalism factor

	Standardized factor loadings	<i>t</i> -test	Probability <i>t</i> -test	R^2	Variance extracted
F: Teacher professionalism ($\rho_c = 0.95$)					0.74
F1 clarity of objectives	0.99			0.98	
F2 value of subject-matter	0.81	19.98	0.001	0.66	
F3 build-up of subject-matter	0.86	19.34	0.001	0.75	
F4 presentation skills	0.85	18.05	0.001	0.72	
F5 (course materials) contribution to understanding the subject	0.94	22.65	0.001	0.88	
F6 course difficulty	0.89	21.29	0.001	0.79	
F7 help of the teacher during the learning process	0.65	11.42	0.001	0.43	

Notes: Fit statistics for confirmatory factor analysis of 7 indicators for 1 construct: $\chi^2_{(30)} = 54.472, p = 0.00$; GFI = 0.98; CFI = 0.97; PNFI = 0.63; NNFI = 0.96; RMSEA = 0.061. We let the error covariances between F3 and F4 and between F4 and F5 correlate.

indeed the existence of such a factor, which accounts for at least 65% for each of seven sub-dimensions.

The fit indices demonstrate a reasonable fit to the data (RMSEA = 0.061, NNFI = 0.96, CFI = 0.97). In our view, the higher-order factor in this study can be seen as a 'Teacher professionalism' factor, as it influences those factors that measure the way a teacher built, organized and executed his/her course. If he/she managed to do this professionally, it will be rewarded by the students as 'good teaching' and thus carry higher ratings on the seven scales in our model. In further analysis, we will use this 'Teacher professionalism' factor to examine the relationship between grading and student ratings and potential influencing factors.

Grading effect and/or other influencing factors?

Concerning the relationships between absolute grade in the course ('examination score') and student ratings of teaching, Cohen (1977) provided some guidelines to interpret the strength of the correlation. He stated that a coefficient of 0.50 is large, 0.30 is moderate and 0.10 is small (Cohen, in Kulik, 2001, p.12). In the second phase of the study, we put student ratings on the seven dimensions of the 'Teacher professionalism' factor and the final examination scores in a Lisrel model, which resulted in a regression coefficient of 0.50 (Model 1 in Table 5); this indeed suggests a strong direct effect of final grade in the course on teacher professionalism.

Table 5. Correlation matrix ratings–grade–overall grade

Model	Independent	Dependent			Test statistics
		Grade	Overall grade	Teacher professionalism	
Model 1	Grade			0.50**	$\chi^2_{(18)} = 41.00$ $p = 0.00$ RMSEA = 0.076 NNFI = 0.96 CFI = 0.97
Model 2	Grade		0.65**	0.36**	$\chi^2_{(24)} = 43.72$ $p = 0.00$ RMSEA = 0.061
	Overall grade			0.22*	NNFI = 0.97 CFI = 0.98
Model 3	Grade		0.61**	0.41**	$\chi^2_{(45)} = 91.63$ $p = 0.00$
	Overall grade			0.22*	RMSEA = 0.068
	Sex	-0.10	0.05		NNFI = 0.94
	Class size	-0.25**	-0.11*		CFI = 0.96
	Course attendance	0.27**	0.15*		

* $p < 0.01$; ** $p < 0.001$.

However, when we brought in the overall grades (students' grades over all courses) and calculated the influence on the teacher professionalism scores, we found a slightly different relationship (Model 2): the direct effect of final grade in the course on the teacher professionalism factor drops to 0.36, which stands for a moderate relationship. It is thus suggested there exists a strong correlation between grade in the course and overall grade (0.65), since the grades the students receive for each course they attend set their overall score at the end of the academic year or semester. But particularly, the direct influence of overall grade on student ratings with regard to one specific course (0.22) suggests an interesting conclusion: it seems to be the case that better students give higher ratings.

This effect still exists in Model 3, where we brought in three background variables, namely class size, course attendance and gender. Interesting parameters here are the negative effect of class size on both final grade in the course and overall grade, and the effect of course attendance on final grade in the course. Students who usually attend the lectures that are given in smaller classes might have better learning outcomes. The effects of gender on final grade in the course and overall grade are not significant.

All three models have a reasonable fit to the data, although the χ^2 test of exact fit is significant each time. However, the better placed root mean square error of approximation, the non-normed fit index and the comparative fit index suggest that the three models provide an acceptable representation of the observed data.

Summary

We can summarize the results of the present study as follows:

1. Confirmatory factor analysis showed the existence of an underlying factor (which we call 'Teacher professionalism') that influences student ratings on seven of 10 strongly validated teacher performance scales. It seems to be the case that teachers who build up and organize their course in a professional and well-considered way receive higher ratings on several domains of the course, since they are considered 'professional teachers' by their students.
2. The results of our study also indicated that there indeed exists a moderate to strong effect of students' grades in a course on student ratings for that course. However, we cannot ignore the overall grade's influence on student ratings: better students (in terms of higher grades over all courses) give higher ratings on teaching effectiveness in a particular course. This conclusion partly refutes the 'grading leniency hypothesis' that argues that teachers who give higher grades will be rewarded with higher student ratings, since the overall grade moderates this relationship. An important topic for further research concerns the question of how the correlation between overall grade and student ratings for a particular course must be interpreted.
3. We also found significant effects of class size and course attendance on final grade in the course, and of class size on overall grades. This suggests that students who

usually attend the lectures that are given in smaller classes might have better learning outcomes.

Discussion

Although these conclusions seem interesting and thus may have implications for further research, we cannot ignore the limitations of the present study. First, this study was limited to a small sample of 222 students at a single university; before generalizing the results, research with larger samples that provide further evidence is required. A second limitation concerns the absence in our models both of the workload and the prior subject interest factors with regard to the relationship between grading and student ratings; other research has shown the important role these play in understanding such correlations (Greenwald & Gillmore, 1997a, 1997b; Marsh & Roche, 2000; Centra, 2003). However, it is our intention to incorporate questions concerning these factors in a further research project on this subject.

Despite these limitations, the results of this study underline the value of student evaluation of teaching effectiveness. It seems to be the case that good (i.e. professional) teachers receive higher ratings, although other factors might still influence these evaluations. Besides, in this study, we found no evidence for the grading leniency hypothesis, ... as the strong relationship between grading in the course and student evaluation of that course is moderated by other variables (i.e. the student's overall score); however, the validity hypothesis receives more support: good ratings probably reflect good learning.

Conclusion

The use of student evaluations of teaching performance has been an important but controversial tool in the improvement of teaching quality during the past few decades. Although student evaluations of teaching are implemented in many faculties, not everyone is convinced of the desirability and utility of such ratings. The present study, however, underlines the value of student evaluations, since we found that students reward good teachers with higher ratings on several scales of teacher performance. We found no support for the 'grading leniency hypothesis', which suggests a problematic biasing factor that could seriously undermine the validity of student evaluations.

References

- Anderson, J. C. & Gerbing, D. W. (1988) Structural equation modeling in practice: a review and recommended two-step approach, *Psychological Bulletin*, 103, 411–423.
- Bagozzi, R. P. & Yi, Y. (1988) On the evaluation of structural equation models, *Academy of Marketing Science*, 16, 74–94.
- Brodie, D. A. (1998) Do students report that easy professors are excellent teachers? *Canadian Journal of Higher Education*, 28(1), 1–20.
- Centra, J. A. (1993) *Reflective faculty evaluation* (San Francisco, CA, Jossey-Bass).

- Centra, J. A. (2003) Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education*, 44(5), 495–518.
- Chen, Y. & Hoshower, L. B. (2003) Student evaluation of teaching effectiveness: an assessment of student perception and motivation, *Assessment and Evaluation in Higher Education*, 28(1), 71–88.
- Chonko, L. B., Tanner, J. F. & Davis, R. (2002) What are they thinking? Students' expectations and self-assessments, *Journal of Education for Business*, May–June, 271–281.
- Cohen, J. (1977) *Statistical power analysis for the behavioral sciences* (New York, Academic Press).
- Feldman, K. A. (1976) Grades and college students, evaluations of their courses and teacher, *Research in Higher Education*, 4, 69–111.
- Feldman, K. A. (1997) Identifying exemplary teachers and teaching. Evidence from student ratings, in: R. Perry & J. Smart (Eds) *Effective teaching in higher education: research and practice* (New York, Agathon).
- Fornell, C. & Larcker, D. F. (1981) Evaluating structural equation models with unobservable variables and measurement error, *Journal of Marketing Research*, 18, 39–50.
- Greenwald, A. G. & Gillmore, G. M. (1997a) Grading leniency is a removable contaminant of student ratings, *American Psychologist*, 52(11), 1209–1217.
- Greenwald, A. G. & Gillmore, G. M. (1997b) No pain, no gain? The importance of measuring course workload in student ratings of instruction, *Journal of Educational Psychology*, 89(4), 743–751.
- Hatcher, L. (1994) *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling* (Cary, SAS Institute).
- Knapper, C. (2001) Broadening our approach to teaching evaluation, *New Directions for Teaching and Learning*, Winter, 88, 3–9.
- Krautmann, A. C. & Sander, W. (1999) Grades and student evaluation of teachers, *Economics of Education Review*, 18, 59–63.
- Kulik, J. A. (2001) Student ratings: validity, utility and controversy, *New Directions for Institutional Research*, 27(5), 9–25.
- Marsh, H. W. (1984) Students, evaluations of university teaching: dimensionality, reliability, validity, potential biases and utility, *Journal of Educational Psychology*, 76(5), 707–754.
- Marsh, H. W. (1987) Students' evaluations of university teaching: research findings, methodological issues, and directions for further research, *International Journal of Educational Research*, 11(3), 253–388.
- Marsh, H. W. & Roche, L. A. (2000) Effects of grading leniency and low workload on students' evaluations of teaching: popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology*, 92(1), 202–228.
- Marsh, H. W. *et al.* (1988) Goodness-of-fit indexes in confirmatory factor analysis: the effect of sample size, *Psychological Bulletin*, 103, 391–410.
- McKeachie, W. J. (1997) Student ratings: the validity of use, *American Psychologist*, 52, 1218–1225.
- Mortelmans, D. & Spooren, P. (2005) *Kwaliteit meten en beoordelen. Eindrapport van de valideringsstudie naar het UA-evaluatieinstrument voor opleidingsonderdelen* (Antwerpen, Universiteit Antwerpen, Faculteit Politieke en Sociale Wetenschappen).
- Penny, A. R. (2003) Changing the agenda for research into students' views about university teaching: four shortcomings of SRT research, *Teaching in Higher Education*, 8(3), 399–411.
- Pike, R. P. (1999) The constant error of the halo in educational outcomes research, *Research in Higher Education*, 40(1), 61–86.
- Shevlin, M., Banyard, P., Davies, M. & Griffiths, M. (2000) The validity of student evaluation of teaching in higher education: love me, love my lectures? *Assessment and Evaluation in Higher Education*, 25(4), 397–405.
- Sproule, R. (2002) The underdetermination of instructor performance by data from the student evaluation of teaching, *Economics of Education Review*, 21, 287–294.

- Theall, M. & Franklin, J. (2001) Looking for bias in all the wrong places: a search for truth or a witch hunt in student ratings of instruction? *New Directions for Institutional Research*, 27(5), 45–56.
- Wachtel, H. K. (1998) Student evaluation of college teaching effectiveness: a brief review, *Assessment and Evaluation in Higher Education*, 23(2), 191–210.
- Wolfer, T. A. & Johnson, M. (2003) Re-evaluating student evaluation of teaching: the teaching evaluation form, *Journal of Social Work Education*, 39(1), 111–120.

Copyright of Educational Studies (Carfax Publishing) is the property of Routledge and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.