

The relationship between professional behaviour grades and tutor performance ratings in problem-based learning

DIANA HJM DOLMANS, SCHELTUS J VAN LUIJK, INEKE HAP WOLFHAGEN & ALBERT JJA SCHERPBIER

PURPOSE We investigated the influence of harsh grading by tutors on tutor performance rating by students.

METHODS A total of 187 tutors assessed students' professional behaviour in tutorial groups. Students rated tutor performance after receiving their grades for professional behaviour. In addition, students were asked to indicate whether they perceived their professional behaviour grades as too positive, adequate or too negative. This was considered to reflect tutors' harshness of grading. Students also rated the quality of the feedback they received from tutors with respect to their grades.

RESULTS Professional behaviour grades that students perceived as too negative, adequate or too positive were associated with tutor performance ratings of 7.4 (SD = 0.9, scale 1–10, $n = 33$), 7.7 (SD = 0.9, scale 1–10, $n = 95$) and 7.5 (SD = 0.8, scale 1–10, $n = 59$), respectively. Harshness of grading did not influence tutor performance ratings significantly. Tutor ratings were predicted more effectively by the quality of the feedback tutors provided on grades than by the harshness of grading.

CONCLUSIONS Tutor performance ratings were not related significantly to harshness of grading. Two explanations can be given: (1) tutor performance ratings were based on rating by groups of students and (2) the percentage of tutors who rated students' professional behaviour as unsatisfactory was low. The strong relationship between tutor performance ratings and the adequacy of the feedback given by tutors

suggests that the tutor performance ratings collected in this study are a valid measure of the quality of their teaching, although, for a full picture of teaching quality, more measures will be needed.

KEYWORDS education, medical, undergraduate; *educational measurement; *interprofessional relations; teaching/ *stands; professional competence/ *standards; students, medical/ *psychology; *problem-based learning; biofeedback (psychology); perception; attitude of health personnel; humans.

Medical Education 2006; **40**: 180–186

doi:10.1111/j.1365-2929.2005.02373.x

INTRODUCTION

Problem-based learning (PBL) is assumed to have positive effects on student learning.¹ However, the expected positive effects of group work do not always materialise. This happens not only in problem-based learning, but also with other educational group formats.² Problems in group work include students who keep up the semblance of hard work while letting others perform the real work. One way of dealing with such problems is to have tutors judge summatively students' individual contributions to the group process.³

In the problem-based curriculum of Maastricht Medical School, the Netherlands, tutors assess students' professional behaviour, including their active participation in group sessions, halfway through (formative) and at the end of modules (summative). During the last few years professional behaviour has come to occupy an increasingly prominent position in medical education. At Maastricht, it has been defined as a combination of 3 aspects: dealing with tasks (participation in discussions in the group),

University of Maastricht, Maastricht, the Netherlands

Correspondence: Dr DHJM Dolmans, University of Maastricht, Department of Educational Development and Research, PO Box 616, 6200 MD Maastricht, the Netherlands.
E-mail: d.dolmans@educ.unimaas.nl

Overview

What is already known on this subject

Research has shown that student ratings are valid and relatively unaffected by biases such as harsh grading.

What this study adds

Tutor performance ratings can be explained more effectively by the quality of tutors' feedback on students' professional behaviour grades than by lower-than-deserved grades or harsh grading, even in an authentic situation where students know their final grades before rating tutor performance.

Suggestions for future research

Further research is needed to investigate the harshness effect in situations where the percentage of tutors giving unsatisfactory grades is much higher.

dealing with others (collaboration) and dealing with oneself (adequate response to feedback and reflective ability). It is graded as satisfactory or unsatisfactory at the end of modules from year 1 onwards. The idea behind early assessment is that feedback and judgement will stimulate students' professional growth from the beginning of their studies throughout their professional life.^{4,5}

There is, however, a danger of bias in the assessment of professional behaviour at Maastricht due to the timing of both this assessment and tutor evaluation by students. In the last tutorial group meeting of every module, students rate anonymously the performance of their tutor who graded their professional behaviour 2–3 days before that. Student ratings of tutor performance are used for staff appraisal interviews, and negative evaluations may have repercussions on promotion decisions.⁶ It is thus vital that these evaluations should be valid and reliable and not vulnerable to bias introduced by students' reactions to harsh or lenient grading by tutors. One of the complaints by tutors is that their evaluation is distorted by low ratings from students whose professional behaviour was graded as unsatisfactory. In their

opinion, ratings reflect tutors' leniency or harshness in judging professional behaviour rather than their real contribution to student learning. In the literature, this phenomenon is referred to as the grading leniency effect. Tutors may receive higher-than-deserved ratings from students because they give higher-than-deserved grades. The opposite of the leniency effect is the harshness effect, i.e. low-grading tutors may receive lower-than-deserved ratings.^{7–9} In the study reported in this paper we examined the harshness effect.

The grading leniency effect has been studied extensively. In one of a series of articles on the validity of student ratings of instruction, published in the *American Psychologist* in 1997, Marsh and Roche concluded that student ratings were relatively unaffected by biases such as grading leniency.^{7,8} In the same series, however, Greenwald and Gillmore concluded that instructors may obtain higher ratings by giving higher grades and that grading leniency had stronger than minor effects on rating.^{9,10} In 1999, Van Os concluded that student ratings were valid and that grading leniency had only minor effects.¹¹ In 2000, Marsh and Roche published an analysis of Greenwald and Gillmore's data published in 1997¹⁰ and a study in which they presented new analyses.¹² They found limited support for the grading leniency hypothesis and concluded that evidence for this effect was hard to find.¹² Their argument that bias was limited was based on the low correlation between students' grades and ratings of instruction (about 0.20) and on the contributions of other valid factors; for instance, the fact that the highest correlation was with learning. In their view, student ratings were multi-dimensional and good teaching produced many desirable outcomes, including better learning and higher expected grades.¹²

Although one might question whether it is necessary to investigate further the leniency effect when it has already been researched extensively, we believe that there are several good reasons for doing so. The first is that the situation in Maastricht offers an excellent opportunity to study this phenomenon in a natural situation. Most studies were conducted in experimental situations where, for example, students were asked to rate teacher performance after being told they had failed a test, irrespective of the real test result. We were able to perform our study in a natural, albeit atypical, situation where students received their real grades a few days before judging their tutors. Usually, students rate tutor performance before they know their grades or test results. A second reason is that we used a different, more

appropriate definition of grading leniency. The grading leniency hypothesis states that instructors who give higher-than-deserved grades are rewarded with higher-than-deserved ratings. This suggests that students' perceptions as to the leniency or harshness of grades might be a better predictor of tutor ratings than the grades themselves. That is why we used both actual grades and the leniency or harshness of these grades as perceived by students. Finally, the unit of analysis in this study, i.e. average tutor ratings across students, is more appropriate than the ratings used in some other studies. Student ratings of tutor performance should not be derived from rating by individual students, but should be based on average tutor ratings across students.

In Marsh and Roche's opinion, a growing body of misleading research appears to fuel myths on student ratings.¹² Some typical problems are: inappropriate definition of bias, neglect of the multi-dimensionality of student ratings, inappropriate use of the student as the unit of analysis instead of class averages, small samples, causal interpretations of correlations and inappropriate experimental manipulations. In the study we conducted, we resolved some of these problems; that is, the situation was authentic, the definition of grading leniency more appropriate and we used average ratings across students.

We hypothesised that lower grades for professional behaviour lead to tutor ratings that are lower than those associated with higher grades for professional behaviour. In addition, we assumed that students who perceive their grades as lower-than-deserved give lower tutor ratings than students who perceive their grades as appropriate. The opposite was expected to apply for students whose grades were higher-than-deserved. We also hypothesised that there is an effect of the quality of the discussions by tutors and students about professional behaviour grades. The idea is that tutor performance ratings are influenced not only by professional behaviour grades, but also by the quality of the feedback students receive, regardless of whether grades are positive or negative. Appropriate feedback is assumed to make a positive contribution towards student learning and this will be reflected in tutor ratings.

We addressed the following research questions: (1) Do professional behaviour grades influence tutor performance ratings by students? (2) What effect do judgements that are perceived as harsh or lenient by students have on tutor performance ratings? (3) How does the way tutors discuss professional behaviour

grades with students impact on tutor performance ratings?

METHOD

Context

The study was conducted at Maastricht Medical School, the Netherlands. We collected data from 5, 4, 1 and 4 6-week modules in years 1, 2, 3 and 4, respectively. Students work in tutorial groups of about 10 students, facilitated by a tutor. Students and tutors meet twice-weekly for 2-hour sessions. It is the task of the tutor to stimulate and facilitate student learning in the group, for instance by asking stimulating questions.

Subjects

A total of 2990 sets of ratings were collected from 849 different students in 341 tutorial groups. Nearly half the 187 tutors included in the study guided 2 tutorial groups within 1 module.

Variables

Students completed a questionnaire in which they rated tutor performance on a scale from 1 to 10 (6 = satisfactory, 10 = excellent). The questionnaire consisted of several items based on theoretical notions concerning effective tutoring. These items represent 5 underlying factors. The reliability and construct validity of this instrument were tested in earlier studies, in which the factors underlying the items were found to have a good fit to the data. The average factor and item score correlated highly with overall tutor performance ratings. In addition, alpha coefficients demonstrated acceptable levels.^{13,14}

Data were collected regarding 6 variables. Tutor performance rating is the first variable (V1) and the second variable (V2) consists of professional behaviour grades given by tutors on a 2-point scale (unsatisfactory or satisfactory).⁴ The content validity of the professional behaviour instrument has been investigated; that is, the 3 factors and underlying items of this instrument are based on theoretical notions regarding professional behaviour. The factors are: (1) how students deal with tasks (participate in group discussions); (2) how students deal with others (collaborate); and (3) how students deal with themselves (respond to feedback and reflect on their performance). Variable 3 concerns tutors' harshness or leniency. Students were asked to indicate whether

their end-of-module grade for professional behaviour was: too positive (1), appropriate (2) or too negative (3) (V3). The 3 remaining variables concern feedback on professional behaviour grades. Students were asked to indicate: (1) whether the professional behaviour grade was discussed halfway through the module: no (0) or yes (1) (V4); whether it was discussed at the end of the module: no (0) or yes (1) (V5); and whether they were satisfied with the quality of the discussion about their grades at end of the module: disagree (1), neutral (2), agree (3) (V6). Variables and rating scales are presented in Appendix 1.

Analysis

The data of 14 modules were pooled and aggregated at tutor level ($n = 187$). Average scores were calculated for all variables. We analysed the data at tutor level (mean ratings across students) and not at student level, because we were interested in the effect of tutors' harshness or leniency on tutor performance ratings. The average scores were categorised. The average professional behaviour grade (V2) was recoded as satisfactory when all students in a group had received satisfactory grades (174 tutors or 93%). In all other cases, when 1 or more students' grades were unsatisfactory, the grade was recoded as unsatisfactory (13 tutors or 7%). For variables 4 and 5 (V4 and V5), which were rated initially on a 2-point scale: no (0) or yes (1), average scores of $= 0.5$ were coded as unsatisfactory and average scores > 0.5 were coded as satisfactory. The ratings on a 3-point scale for variables 3 and 6 (V3 and V6) were categorised as too positive (> 2), adequate ($= 2$) or too negative (< 2) (V3) and as disagree (< 2), neutral ($= 2$) or agree (> 2) (V6), respectively. Analysis of variance (ANOVA) and regression analysis were performed to analyse the data.

RESULTS

Student response varied between 76% and 97% per module, with an average of 87%. The average overall tutor performance rating was 7.6 (SD = 0.8,

$n = 187$). Of the tutors, 7% graded the professional behaviour of 1 or more students as unsatisfactory and 93% of the tutors gave satisfactory grades to all students (Table 1). Of the students, 31% indicated that their final professional behaviour grade was too positive, 51% indicated that it was adequate and 18% thought their grade was too negative (Table 2). Of the students, 13% indicated that their tutors had discussed professional behaviour grades halfway through the module and 87% indicated that their tutors had not done so. The corresponding percentages at the end of modules were 14% and 86%. Finally, the quality of the discussion of professional behaviour grades was regarded as satisfactory, neutral and unsatisfactory by 72%, 19% and 10% of the students, respectively (Table 2).

The ANOVA analysis that was conducted to investigate the first hypothesis indicated that average tutor performance ratings did not differ significantly (see Table 1). An unsatisfactory grade corresponded with a tutor performance rating of 7.4 (SD 0.7). A satisfactory grade corresponded with a tutor performance rating of 7.6 (SD 0.9) ($P = 0.498$, NS). This implies that the first hypothesis was not confirmed by the data; that is, unsatisfactory professional behaviour grades did not correspond with lower tutor performance ratings.

The second hypothesis, that tutors' harshness or leniency (or the match between professional behaviour grades and students' perceptions of the quality of their professional behaviour) influences tutor performance ratings, was also not confirmed by the data. As can be seen in Table 2, tutor performance ratings associated with too positive, adequate and too negative grades as perceived by students are 7.5 (SD = 0.8), 7.7 (SD = 0.9) and 7.4 (SD = 0.9), respectively. The differences between these ratings are not significant ($P = 0.242$).

The only hypothesis that was confirmed was that the quality of feedback on professional behaviour grades is related to tutor performance ratings. The results demonstrate that the average tutor performance rating was significantly higher if professional

Table 1 Results of the analysis of variance examining the effects of professional behaviour grades (V2) on tutor performance ratings (V1) (scale 1–10)

Professional behaviour grade	Mean tutor rating (SD)	No. of respondents	F-value	P-value
Unsatisfactory	7.4 (0.7)	13 (7%)		
Satisfactory	7.6 (0.9)	174 (93%)	0.461	0.498

Table 2 Results of an analysis of variance examining the effects of several variables (V3–V6) on tutor performance ratings (V1) (scale 1–10)

Variable	Mean tutor rating (SD)	No. of respondents (%)	F-value	P-value
V3 The final grade for my professional behaviour given by the tutor at the end of the module was				
Too positive	7.5 (0.8)	59 (31%)		
Adequate	7.7 (0.9)	95 (51%)		
Too negative	7.4 (0.9)	33 (18%)	1.43	0.242
V4 The professional behaviour grade was discussed halfway through the module				
No	7.2 (1.4)	24 (13%)		
Yes	7.6 (0.7)	163 (87%)	6.45	0.012
V5 The professional behaviour grade was discussed at the end of the module				
No	7.0 (1.2)	26 (14%)		
Yes	7.7 (0.7)	161 (86%)	13.9	0.000
V6 I am satisfied with the way the tutor discussed the professional behaviour grade with me at the end of the module				
Disagree	6.7 (1.3)	18 (10%)		
Neutral	7.0 (0.8)	36 (19%)		
Agree	7.9 (0.6)	133 (72%)	31.5	0.000

behaviour grades were discussed with students halfway through and at the end of modules (Table 2). The results also demonstrate significantly higher tutor performance ratings if students are satisfied with the discussion of their grades with tutors (7.9, SD 0.6) compared to a situation where students are not satisfied with the discussion (6.7, SD 1.3). A regression analysis was performed using tutor performance rating (V1) as dependent variable and professional behaviour grade (V2), the quality of the discussion about grades (V6) and harshness or leniency as perceived by students (V3) as independent variables. Table 3 shows the correlations between the variables. Only 1 coefficient is significant at the 0.01 level. There is a strong positive correlation between the quality of the discussion and tutor performance rating (0.60). A linear regression analysis stepwise method ($n = 186$), with the same 3 independent variables, showed that the regression model used only 1 independent variable to explain tutor performance rating (V1), namely student satisfaction with the way their grades were discussed (V6).

The standardised beta weight R was 0.596 (significance 000) and the explained variance was 36%.

DISCUSSION AND CONCLUSION

This study investigated the relationship between students' professional behaviour grades and tutor performance ratings. The results indicate that there is no significant association between high professional behaviour grades and high tutor performance ratings nor between low grades and low tutor ratings. Thus, the hypothesis that a low professional behaviour grade leads to a lower tutor performance rating was not confirmed by the data.

Tutor performance rating was not found to be affected by harshness or leniency of grading as perceived by students. Students' perceptions about tutors' harshness or leniency can be regarded as a stronger measure of the harshness or leniency effect than actual professional behaviour grades,

Table 3 Pearson correlations

	Tutor performance rating (V1)	Professional behaviour grade (V2)	Student satisfaction with feedback on grade (V6)
Tutor performance rating (V1)			
Professional behaviour grade (V2)	0.04		
Student satisfaction with feedback on grade (V6)	0.60**	0.15*	
Student opinion about leniency or harshness of grade (V3)	- 0.05	- 0.18*	- 0.11

*Significant at 0.05 level, **significant at 0.01 level.

because students' perceptions of harshness or leniency depend on whether they see their grades as an inadequate reflection of their performance. The evidence for the absence of a harshness or leniency effect is strengthened by the fact that this study was performed in a non-experimental, authentic situation where students rated tutor performance after receiving their grades from these tutors. Two explanations for the absence of the harshness or leniency effect can be given. First, the effect of individual student ratings was very small, because in the calculations we used the ratings of groups of at least 6 students. Secondly, the percentage of tutors who graded 1 or more students' professional behaviour as unsatisfactory was low (7%, 13 tutors). It should be noted that the professional behaviour grades used in this study have limited discriminating value and the finding that 31% of the students perceived their grades as too positive indicates that this instrument warrants further study.

The results indicate that the quality of tutor feedback as perceived by students is a better predictor of tutor performance ratings than is the harshness or leniency effect. This suggests that tutor performance ratings were not influenced by harsh or lenient grading. However, this finding can also be described as a halo effect.

The findings of this study imply that at Maastricht Medical School tutor performance ratings are not affected by either the leniency or the harshness effect, even in an authentic situation where students know their final grade before rating tutor performance. Further research will need to investigate the harshness effect in situations where the percentage of tutors who give unsatisfactory grades is much higher. In addition, further research will need to examine why 31% of the students deemed their professional behaviour grades to be too positive. Tutors may have difficulty grading students' professional behaviour because they are not good at facilitating this type of behaviour. The tutor ratings used in this study are part of the medical school's regular quality improvement process. They are used in the reward system for teacher performance, which is included in the medical school's faculty development plan. The strong relationship between tutor performance ratings and the adequacy of the feedback given by tutors suggests that the tutor performance ratings collected in this study are a valid measure of the quality of their teaching. The results demonstrate that these tutor ratings are indeed valid indicators of tutor performance and that there is no evidence of a grading

leniency effect. Nevertheless, readers should keep in mind that effective teaching is multi-dimensional and that student ratings of tutor performance should not be the only source of information and should not be over-interpreted.

Contributors: DD designed the study, analysed the data and wrote the paper. IW was involved in the design of the study. SL was responsible for part of the data collection. All authors contributed to the writing of the paper.

Acknowledgements: we would like to thank Diana Riksen for setting up the data-set. Thanks to Mereke Gorsira for revising the English.

Funding: none.

Conflicts of interest: none.

Ethical approval: not required.

REFERENCES

- 1 Norman GR, Schmidt HG. The psychological basis of PBL: a review of the evidence. *Acad Med* 1992;**67**(9):557–65.
- 2 Michaelsen LK, Fink LD, Black RH. What every faculty developer needs to know about learning groups. In: Richlin L, ed. *To Improve the Academy*, vol. 15. Stillwater, OK: New Forums Press 1996;31–57.
- 3 Dolmans DHJM, Wolfhagen HAP, van der Vleuten CPM, Wijnen WHFW. Solving problems with group work in PBL: hold on to the philosophy. *Med Educ* 2001;**35**:884–9.
- 4 Van Luijk SJ, Smeets JGE, Smits J, Wolfhagen I, Perquin MLF. Assessing professional behaviour and the role of academic advice at the Maastricht Medical School. *Med Teach* 2000;**22**(2):168–72.
- 5 Arnold L. Assessing professional behaviour: yesterday, today and tomorrow. *Acad Med* 2002;**77**(6):502–15.
- 6 Dolmans DHJM, Wolfhagen HAP, Schmidt HG, van der Vleuten CPM. A rating scale for tutor evaluation in a problem-based curriculum: validity and reliability. *Med Educ* 1994;**28**:550–8.
- 7 Greenwald AG. Validity concerns and usefulness of student ratings of instruction. *Am Psychol* 1997;**52**(11):1182–6.
- 8 Marsh HW, Roche LA. Making students' evaluations of teaching effectiveness effective. *Am Psychol* 1997;**52**(11):1187–97.
- 9 Greenwald AG, Gillmore GM. Grading leniency is a removable contaminant of student ratings. *Am Psychol* 1997;**52**(11):1209–17.
- 10 Greenwald AG, Gillmore GM. No pain, no gain? The importance of measuring course workload in student ratings of instruction. *J Educ Psychol* 1997;**89**(4):743–51.
- 11 Van Os W. Bruikbaarheid en Effectiviteit van Studentoordelen over het Onderwijs [Usability and effectiveness of student ratings of education]. Thesis,

- University of Amsterdam, the Netherlands. Enschede: Ipskamp 1999.
- 12 Marsh HW, Roche LA. Effects of grading leniency and low workload on students' evaluations of teaching. popular myths, bias, validity or innocent bystanders? *J Educ Psychol* 2000;**92**(1):202–28.
- 13 Dolmans D, Wolfhagen HAP, Scherpbier AJJA, van der Vleuten CPM. Development of an instrument to evaluate the effectiveness of teachers in guiding small groups. *Higher Education* 2003;**46**:431–46.
- 14 Dolmans D, Ginns P. A short questionnaire to evaluate the effectiveness of tutors in PBL. validity and reliability. *Med Teach* 2005;**27**(6):534–538.

Received 11 January 2005; editorial comments to authors 8 February 2005; accepted for publication 27 June 2005

Appendix 1 The variables used in this study and the corresponding rating scales

Variable	Scale
V1 Tutor performance rating (by students)	1–10
V2 Professional behaviour grade (by tutors)	Unsatisfactory (0), satisfactory (1)
V3 The final grade for professional behaviour given by the tutor at the end of the module was:*	Too positive (1), adequate (2), too negative (3)
V4 The professional behaviour grade was discussed with me halfway through the module	No (0), yes (1)
V5 The professional behaviour grade was discussed with me at the end of the module	No (0), yes (1)
V6 I am satisfied with the way in which the tutor discussed my professional behaviour grade with me at the end of the module	Disagree (1), neutral (2), agree (3)

*This variable (V3) is referred to as tutors' harshness or leniency.