Student Evaluations of Teaching Assistants in a First Year Physics Laboratory : Is there a Grade Dependence?

by Z. Hazari and T. Key

In this study we examined the dependency of 630 student evaluations on grades assigned to students in a first year physics laboratory taught by 38 graduate teaching assistants (TAs). Although the overall student evaluations of the TAs were dependent on their assigned grades, a more detailed examination revealed that only questions related to the TA's fairness and motivational influence were dependent on grades whereas questions related to the usefulness of the TA's assistance were not.

A more detailed examination revealed that only questions related to the Teaching Assistant's fairness and motivational influence were dependent on grades whereas questions related to the usefulness of the TA's assistance were not.

INTRODUCTION

Student evaluations of instructors and classes provide valuable information about the quality of instruction. These are especially useful in the case of laboratory courses, which are highly interactive and often graded subjectively by the instructor. However, there are concerns within the educational community that student evaluations are not a valid form of assessment of teaching or teachers and are not reliable guides for the reform of teaching practices. This concern stems from the fact that student evaluations are often directly related to student grades. For example, if a student has a high grade in a course, he or she might be inclined to evaluate the teacher highly regardless of actual teaching performance. This would make the evaluation a poor measure of the quality of teaching - although a good measure of the student's approval of his or her grade!

Many studies have been done to determine the value of student evaluations. However, the results of these studies are frequently contradictory. Eiszler [1] finds a predictive relationship between student ratings of teaching and expected grades and Murkison and Stapleton [2] report that students expected high grades from teachers rated highly. Krautmann and Sander's [3] results indicate that instructors can "buy" better evaluations through more lenient grading. Similarly, Brodie [4] reports that professors who assigned the highest grades for the least studying received the highest evaluations. On the other hand, Gigliotti and Buchtel's [5] results support the validity of student evaluations by indicating minimal self-serving and grade bias, and Gramlich and Greenlee [6] find little relationship between the rating of instructors and the grades received by students. Marsh and Roche [7] present two studies that refute the hypothesis that student evaluations of teaching are biased by grading leniency. In addition, they report that one of the typical problems with the studies that find student evaluations of teaching unreliable is their "neglect of the multidimensionality" of student evaluations. Theall and Franklin [8] affirm that there are many aspects of an instructor's teaching that students can rate accurately as well as some they cannot. For example, students can assess the clarity of an instructor's explanations and the instructor's helpfulness but they cannot accurately rate the instructor's knowledge of the subject. Thus, they conclude that having multiple sources of data and asking questions that students can legitimately answer is necessary for effective evaluation. Arreola [9], after a survey of the literature, concludes that '…the belief that student ratings are highly correlated with their grades is not supported by the literature'.

Clearly, the question of whether students can evaluate their instructors independent of the grades they receive is not straightforward. In this study, we investigate the relationship between students' evaluation of their TA and the grades assigned to students by their TA in a large first year introductory physics laboratory [1]. It is important to understand this relationship in order to minimize the effect of grade bias when designing meaningful student evaluations of teaching.

OVERVIEW

Teaching assistants (TAs) in the first year undergraduate physics laboratory at the University of Toronto are responsible for a large portion of the laboratory grading as part of their teaching duties. The majority of TAs are newly appointed each year and their teaching has not previously been evaluated. Their duties include supervising and guiding the lab work of undergraduate student in groups of 10 to 18 and grading their written and practical work. In the 1999-2000 session, the grades assigned by the TAs to their students con-

[1] Further information on the first year physics laboratory may be obtained at www.upscale.utoronto.ca/PHY110_138Lab.html.

Zahra Hazari <zhazari@oise.utoronto.ca>, Ontario Institute for Studies in Education, Univ. of Toronto, 252 Bloor Street West, Toronto, ON, M5S 1V6; Tony Key <key@physics.utoronto.ca>, Department of Physics, Univ. of Toronto, 60 St. George St., Toronto, ON, M5S 1A7

stituted 70% of the final laboratory grade; the remaining 30% of the grade was determined by a laboratory test, common to all students. The grade provided by the TA was composed of the grades students received on their written work in laboratory notebooks (35%) and formal reports (20%), and their performance during the labs (15%). These grades are subjective, since they depend on the TA's judgement of students' written work and observed performance. When marking subjective lab work, TAs were asked to keep their group averages between 65 and 75 percent.

In March 2000, we broadly administered an evaluation of TAs to students enrolled in the first year laboratory at the University of Toronto. This was a challenging task given that the first year laboratory had enrollments of over a thousand students and there were up to 150 students spread out in four lab rooms and a computer lab during any given lab session. Students in a single TA's group were often found spread out in different locations depending on the lab they were doing or the stage of the lab work. Thus, to simplify our task, we administered the student evaluations in the lecture sections of the three first year courses under study; Basic Physics (non-calculus based course), Physics for the Life Sciences, and Foundations of Physics (for majors). All three courses shared the same laboratory. We also circulated through the lab rooms during lab sessions and asked students who had not already completed the evaluation to do so. Due to the complexity of this task, our limited funding, and other evaluation priorities of the course coordinators, we were unable to repeat this evaluation in subsequent years. However, such cross-sectional time data is common in large-scale studies [10].

Students evaluated their TAs on a 5-point Likert scale from "Very poor" (score one) to "Very good" (score five) on the following measures:

- Fairness in grading
- Availability for assistance in lab
- Availability for assistance outside the lab
- Usefulness of assistance
- Communication skills
- Friendliness and approachability
- Fairness to students - no favorites
- Energy and enthusiasm
- Influence on attitudes to physics
- Knowledge and understanding of physics
- Ability to stimulate to think
- Ability to inspire to do best work
- Overall rating

The validity [1] of the student evaluation was determined through student interviews and it was accordingly revised prior to administration. We collected 630 student evaluations from the students of 38 of the 44 TAs teaching in the first year laboratory. There were 79, 486, and 65 evaluations from Basic Physics, Physics for the Life Sciences, and Foundations of Physics respectively. Since each item on the evaluation could be ranked from one to five and there were 13 items, the total score range was from 13 to 65. We also had records of all the student grades in the laboratory.

RESULTS

Since students knew the subjective grade they were assigned by their TA when they filled in the evaluation forms, the student evaluations were taken to be the dependent variable and the grades were treated as the independent variable. We were interested in investigating whether the grades received by the students could predict their evaluations. Had TAs received their evaluations prior to assigning grades, we might ask whether student evaluations predict the grades assigned. However, this bias is removed since instructors at University of Toronto do not see their evaluations until well after submitting grades.

We first performed a multiple linear regression to determine if the subjective grade assigned by the TA predicted the evaluation she/he received. A linear regression is a linear model of the form

$$Y_i = B_0 + B_1 X_{1i} + B_2 X_{2i} + \dots B_p X_{pi} + E_i$$

where: Y is the outcome variable; X_1, X_2, ... X_p are the predictor variables; B_0, B_1, B_2, ...B_p are the fitted values or parameter estimates; E is the error or model deviation; and i = 1, 2, ...n for n observations. In our case, the outcome variable is the student evaluation score (*EvalTotal*) received by the TA and the focus predictor is the subjective grade assigned by the TA (*TAmark*). We used control predictor variables to account for the effect of the different courses; Basic Physics and Foundations of Physics were put into the regression as dichotomous dummy variables (0=not enrolled, 1=enrolled) being compared to the largest course, Physics for the Life Sciences (All three courses were not put into the model as this would over-specify the model. By leaving the largest course out, the results serve to compare the two other courses to it). This control was necessary in order to alleviate the effect of the different courses (i.e. types of students taking the courses) on the evaluations. We also controlled for the students' lab-test grade, which is an objective measure of performance, to alleviate the argument that students with high test performance always give better evaluations than students with lower test performance regardless of actual teaching.

The results of the multiple regression are summarized in Table 1. The significance level (p) indicates the probability that the effect of the predictor on the outcome is due to chance. Thus, if $p<0.05$, then there is less than a five percent probability that the effect is due to chance (i.e. greater than 95% probability of an actual effect). This is the generally accepted minimum level required to deem an effect "significant". The R^2 value indicates the fraction of the variance (spread; standard deviation squared) in the dependent variable that is explained by the model consisting of the chosen predictors. The results in Table 1 indicate that even after controlling for the course type and the student performance level, the *TAmark* is still a significant predictor of *EvalTotal* at the $p<0.05$ level. However, both course type and student performance level are non-significant and the model accounts for very little (2%) of the *EvalTotal* variance. This is not sur-

[1] Whether the evaluation measured what it was intended to measure, i.e. whether students understood the same meaning as the researchers and responded correspondingly.

prising since there are many other variables not included in the model that might influence student evaluations of their TA such as actual teacher characteristics (knowledge, gender, race, first language, etc.), learner characteristics (amount learned, gender, race, socio-economic background, etc.), and course characteristics (pedagogical materials, content, workload, etc.). Our goal was only to discover the predictive ability of grades and not to develop a comprehensive model for EvalTotal.

We examined the data further by hypothesizing that there may be some student evaluation items/criteria that are independent of grades as cited in the literature [7,8]. To that end, we performed an exploratory factor analysis (EFA). An EFA is a statistical procedure that detects trends in relationships between variables based on correlations without the presupposition of a hypothesis. It can be used to reduce the number of variables and/or to detect structure in the relationships between variables, thus, classifying them. In our case, the EFA was used to detect similar trends in the items of the student evaluation, grouping them into factors. When we required a multiple factor solution from this analysis, three distinct logical factors emerged. As summarized in Table 2, the questions in each of the groups formed by the EFA have certain similarities. Based on the types of questions that are thus factored together, we named the three groups of questions *Assistance, Fairness, and Influence*, for convenience of discussion.

TABLE 1
Modeling EvalTotal with control predictors and *TAmark* (N=630)

Predictors	B	Std Err	Sig (p)
Intercept (B_0)	36.96	4.67	***
Course Control			
(Basic Phy – B_1)	-1.65	1.38	NS
(Foundations of Phy – B_2)	2.47	1.51	NS
Performance Control			
(Lab-test – B_3)	-0.02	0.02	NS
Tamark (B_4)	**0.16**	**0.06**	*
R^2		0.02	

NS – Not Significant * p<0.05 ** p<0.01 ***p<0.001

The first factor, *Assistance*, groups all the items related to the quality of the assistance given by the TA, including the affective dimension of 'energy and enthusiasm'. The overall rating of the TA, the last item on the student evaluation, also falls into the factor of *Assistance*. It is clear that *Assistance* is the most closely linked factor to the quality of teaching done by the TA. The maximum value for the *Assistance* score is 35 since there are seven items in this factor worth a maximum of five each. The second factor, *Fairness*, includes three items; two that relate to the TA's fairness (in grading and in their interactions with students in lab) and one based on the students' perception of the TA's approachability and friendliness. The last factor, *Influence*, includes three items that judge the TA's influence on students' attitudes to physics and to the laboratory work. The maximum value for both the *Fairness* and *Influence* scores is 15.

TABLE 2
Student Evaluation Factors

Factor 1: *Assistance*	Factor 2: *Fairness*	Factor 3: *Influence*
•• Availability for assistance in lab •• Availability for assistance outside lab •• Usefulness of assistance •• Communication skills •• Energy and enthusiasm •• Knowledge and understanding of physics •• Overall rating	•• Fairness in grading •• Friendliness and approachability •• Fairness to students – no favorites	•• Influence on attitudes to physics •• Ability to stimulate to think •• Ability to inspire to do best work

We then examined the predictive relationship of grades on the three evaluation factors using regression analyses with the same predictor variables as before. The results are summarized in Table 3. The course type and student performance level were not significant except that TAs in the Foundations of Physics course were evaluated significantly higher on *Fairness* than those in the Physics for the Life Sciences course but only at the p<0.05 level. Although *Fairness* and *Influence* are both significantly dependent on *TAmark*, *Fairness* is much more so (p<0.001). The model also explains more of the variance in *Fairness* (4.4%) than the variance in *Assistance* and *Influence* (1% and 2% respectively). Again, this is not surprising given that there were many other unmeasured variables that influence evaluations. The most surprising and notable result, however, is that the largest factor, *Assistance*, is not dependent on grades.

TABLE 3
Modeling Assistance, Fairness, and Influence with control predictors and *TAmark* (N=630)

	Models								
	Assistance			***Fairness***			***Influence***		
Predictors	**B**	**Std Err**	**Sig**	**B**	**Std Err**	**Sig**	**B**	**Std Err**	**Sig**
Intercept (B_0)	22.72	2.71	***	7.25	1.07	***	6.98	4.09	***
Course Control									
(Basic Phy – B_1)	-0.57	0.80	NS	-0.44	0.32	NS	-0.63	0.35	NS
(Foundations of Phy – B_2)	1.39	0.87	NS	0.75	0.34	*	0.33	0.38	NS
Performance Control									
(Lab-test – B_3)	-0.01	0.01	NS	-0.01	0.01	NS	0.00	0.00	NS
Tamark (B_4)	**0.05**	**0.04**	**NS**	**0.07**	**0.01**	***	**0.05**	**0.02**	**
R^2	0.01			0.044			0.02		

NS – Not Significant * p<0.05 ** p<0.01 ***p<0.001

DISCUSSION

There is some evidence that students tend to evaluate more highly those teachers from whom they receive higher grades [1,2,3,4]. Indeed we observed such an effect; however, it is not a simple one. The regression we performed to discover the predictive relationship of the subjective grades TAs assigned to students (*TAmark*) on the TA Total Evaluation scores (*EvalTotal*) was significant for all 38 TAs. This result was significant despite controlling for course type and student objective performance level. The positive *TAmark* coefficient (B) within the regression indicates that the evaluation scores increase with the grades. A clearer understanding emerges from a factor analysis that provides three categories of the evaluation (*Assistance, Fairness*, and *Influence*).

The regression of the *TAmark* with the *Fairness* category is highly significant and positive implying that the *TAmark* is a good predictor of the *Fairness* evaluation items. Although attribution of motive is impossible, we suggest that this might imply that students tend to judge their TA as acting fairly and influencing them as long as they give the student a good grade - an understandable reaction. The regression with *Influence* is also significant, though at a lower level. Again, it is reasonable to expect that students who do better are actually the ones who have been influenced and inspired to put their best foot forward. The most striking result is that the largest factor extracted from the student evaluations, *Assistance*, was not dependent on grades. Thus, it appears that students can discriminate between their TA's effectiveness in providing assistance in the laboratory from the grade that their TA assigns them. This final result is in good agreement with the research that counters the belief that student ratings are highly correlated with their grades. In addition, the varying levels of dependence of our evaluation factors on grades support the argument made by Marsh and Roche [7] regarding the multidimensionality of student evaluations of teaching.

CONCLUSION

In order to be useful, student evaluations should be independent of the grades students receive. In the student evaluations of TAs that we administered we identify an *Assistance* factor that is relevant to teaching quality. This factor provides a good measure of the TA's ability to be helpful and to provide useful assistance independently of the grades students received from these TAs. The grades students received are correlated with the evaluation through the factors related to the TA's fairness and motivational influence. Thus, it appears that students can evaluate important attributes of their teachers in a physics laboratory, as long as the evaluation asks questions directly relating to the quality of the assistance or the instruction provided. Our results shed some light on the differing findings of research on the influence of student grades on student evaluations of their teachers and suggest that grade bias might be avoided in student evaluations of teaching if the questions asked are chosen with care.

ACKNOWLEDGEMENTS

We wish to thank the Department of Physics for their financial support and cooperation in this research. Z. Hazari would like to thank the Social Sciences and Humanities Research Council of Canada for their support.

REFERENCES

1. C. Eiszler, "College Students' Evaluations of Teaching and Grade Inflation", *Research in Higher Education*, **43**, 483-501 (2002).
2. G. Murkison and R. Stapleton, "Optimizing the Fairness of Student Evaluations: A Study of Correlations between Instructor Excellence, Study Production, Learning Production, and Expected Grades", *Journal of Management Education*, **25**, 269-291 (2001).
3. A. Krautmann and W. Sander, "Grades and Student Evaluations of Teachers", *Economics of Education Review*, **18**, 59-63 (1999).
4. D. Brodie, "Do students report that easy professors are excellent teachers?", *Canadian Journal of Higher Education*, **28**, 1-20 (1998).
5. R. Gigliotti and F. Buchtel, "Attributional Bias and Course Evaluations", *Journal of Educational Psychology*, **82**, 341-351 (1990).
6. E. Gramlich and G. Greenlee, "Measuring Teaching Performance", *Journal of Economic Education*, **24**, 3-13 (1993).
7. H. Marsh and L. Roche, "Effects of Grading Leniency and Low Workload on Student Evaluations of Teaching: Popular Myth, Bias, Validity, or Innocent Bystanders", *Journal of Educational Psychology*, **92**, 202-208 (2000).
8. M. Theall and J. Franklin, "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*, **109**, 45-57 (2001).
9. R.A. Arreola, *Developing a Comprehensive Faculty Evaluation System: A Handbook for College Faculty and Administrators on Designing and Operating a Comprehensive Faculty Evaluation System (2nd ed.)*, Anker Publishing Company, 2000.
10. See: P. M. Sadler and R. H. Tai, "Success in Introductory College Physics: The Role of High School Preparation", *Science Education*, **85**, 111-136 (2001); L. G. Ortiz, P. Heron, and P. S. Shaffer, "Student understanding of static equilibrium: Predicting and accounting for balancing", *American Journal of Physics*, **73**, 545-553 (2005); and N. Nguyen and D. E. Meltzer, "Initial understanding of vector concepts among students in introductory physics courses", *American Journal of Physics*, **71**, 630-638 (2003).