Assessment & Evaluation in Higher Education
Vol. 30, No. 6, December 2005, pp. 581–592

Online versus traditional teaching evaluation: mode can matter

Eyal Gamliel* and Liema Davidovitz
Ruppin Academic Center, Israel

Using an experimental mixed design, this study compared the traditional paper-and-pencil method for evaluating teaching with the online method. Replicating previous findings, the comparison revealed similar evaluation means of the two methods. However, the stability of teaching evaluations using paper-and-pencil twice was substantially higher than the corresponding stability using different methods—online and paper-and-pencil. One possible explanation for this finding is the different visual presentation of the scales: a typical form of the paper-and-pencil method presents the scale horizontally, enabling the subjects to examine the profile of their answers that might result in an artificially lower variability of the evaluations. In contrast, an electronic answering form can abolish this artificial answering effect.

Introduction

Online student evaluation of teaching has become an established practice at many institutions of higher education (Thorpe, 2002). Among the major advantages of online student evaluation of teaching over the traditional paper-and-pencil method are economic gains (Hmieleski & Champagne, 2000; Cummings *et al.*, 2001; Johnson, 2002), the efficiency and precision of the evaluation procedure (Kelly & Marsh, 1999; Nulty, 2001; McGourty *et al.*, 2002), dealing with students' answers to open questions (Layne *et al.*, 1999; Goodman & Campbell, 1999; Johnson, 2002) and several research advantages (Klassen & Smith, 2002). However, online evaluation of teaching has to deal with considerable challenges before it can be effectively applied. Among the major problems of online evaluation of teaching are the relatively low response rate (see Ha *et al.*, 1998; Baum *et al.*, 2001) and the anonymity of the students (Cummings *et al.*, 2001; Dommeyer *et al.*, 2002).

One of the main concerns of using the online method for teaching evaluation is the potential problem of non-response bias that might result from the relatively low

*Corresponding author. Behavioral Sciences Department, Ruppin Academic Center, Emek Hefer 40250, Israel. Email: eyalg@ruppin.ac.il

ISSN 0260-2938 (print)/ISSN 1469-297X (online)/05/060581–12

DOI: 10.1080/02602930500260647

response rates (see Sax *et al.*, 2003). A major aspect of this bias is that online teaching evaluation results might be substantially different from the respective results of the traditional method. Several studies compared the two methods' teaching evaluation results, usually finding similar averages for most measures (Ha & Marsh, 1998; Layne *et al.*, 1999; Thorpe, 2002). Similar results of online and paper-and-pencil methods also characterize surveys collecting data other than teaching evaluation (Carini *et al.*, 2003).

The conclusion from current research is that when the major disadvantages of the online evaluation of teaching are overcome the mode probably does not bias the means of teaching evaluation. However, the above studies compared the two methods using a between-subjects design—students were assigned to either one of two groups that evaluated teaching by online or paper-and-pencil methods. Thus, similar means of the two methods can reflect various within-subjects interpretations. For example, similar means of the two methods can reflect the existence of two subgroups that are influenced in opposite directions by the mode, so that although significant differences between the two methods exist within the subjects, these differences are obscured by the means. Alternatively, similar means might reflect relatively small differences between the two methods within most subjects. While the latter might be the intuitive interpretation of the similar between-modes means, the former represent quite a different interpretation that calls for additional considerations in using either the traditional paper-and-pencil method or the online method. Thus, while previous examination of the teaching evaluation's means of the two methods revealed that 'mode probably does not matter' the question 'does mode matter?' is still left unanswered. The aim of this study is to examine this question.

The preferred design for answering this question is a mixed-design in which the within-subjects differences are compared between groups that vary in the mode of evaluation. The mixed design enables to randomly assign students to an experimental condition in which they evaluate teaching twice using similar methods of evaluation or using different methods of evaluation. The difference score between the two evaluations of each student can be used as the unit of analysis for examining the possible effect of the evaluation mode on an individual level.

Thus, the question 'does mode matter?' answered in this study on an individual level using a mixed design is different from the question that sounds the same but was answered previously by between-subjects design regarding the evaluation means.

Method

Sample

The sample included 198 undergraduate students of the Business Management department in a higher education institution. All the students belong to an executive program designed for students with managerial experience. About 75% of the sample was men. The average age was 33 with a standard deviation of 7.5. The subjects were in their second to fifth trimester and studied in nine different classes.

The subjects agreed to participate in the study upon our request. Teaching evaluation is routinely conducted, so our request to evaluate teaching was not out of the ordinary. The students were informed, however, that their evaluations were part of a research.

In order to motivate the students the experimenter promised to draw a reward in each class.

Each student was asked to evaluate the teaching in three courses taught by different teachers. We selected teachers whose previous evaluations were average, in order to maximize the variability of the evaluations. All students evaluated the same courses twice with two weeks intervals (henceforth: measurement one and measurement two).

Instruments

The instrument used in this study was a questionnaire asking the students to evaluate various aspects of the teacher and the teaching. The questionnaire included seven closed items of which two were global items (Question 5 and 6) inquiring about the overall evaluation of the teacher and the course (the questionnaire is presented in Appendix 1). The subjects were asked to evaluate teaching on a 15-point scale, in which 1 is very low and 15 is very high. The scale was grouped into five broader categories, each containing three points. The subjects were asked to give their age and gender.

The questionnaire used in this research was an abridged version of the standard questionnaire. This shortened version was used because the subjects were asked to evaluate three courses in each of the two measurements.

The questionnaire was adapted to the two methods: paper-and-pencil and online. The same questions were used in both methods, although the two methods differ in their visual presentation of the scale used to evaluate teaching: The scale was presented horizontally in the paper-and-pencil format, whereas in the online version a vertical scale opened for each question in a separate window that closed after the subject marked the evaluation.

Design and procedure

Each of the nine classes evaluated the teaching twice within 14-day interval. Out of the nine classes selected for this study, six classes constitute three cohorts of students that for technical reasons were divided into two classes. Each of the two classes studied the same courses with the same teachers. One of these three couples was randomly assigned to a group that evaluated teaching using the traditional method in the first measurement and using the online method in the second measurement. The other three classes were assigned to evaluate teaching using the online method in the first measurement and using the paper-and-pencil method in the second measurement. The subjects in the parallel classes were asked to evaluate the same teachers. These subjects will be referred to henceforth as the experimental group. The remaining

three classes were the control group that evaluated teaching using the paper-and-pencil method twice (henceforth: the control group). The gender and age distributions were similar in all three groups.

The study was conducted during class time: teaching evaluations using paper-and-pencil were conducted during a regular lesson, while the online teaching evaluations were conducted during computer laboratory lessons. The above procedure was preferred over the one in which online evaluation is performed during the student's free time because of the different experimental and control groups that might result from such a procedure: following Eckel and Grossman (2000) students evaluating teaching in class might be referred to as pseudo-volunteers while students evaluating teaching on their free time might be referred to as volunteers. Pseudo-volunteers and volunteers are two distinct groups that might provide different evaluations. The procedure used in this study overcomes the possible problem that might be encountered using online evaluation, in which there is no control over how and with whom the evaluations are completed (Theall, 2000).

Each student received an arbitrary code. The code enabled us to match the two evaluations and ensured the confidentiality of the evaluation. In addition, the codes enabled experimental group subjects to gain access to the online evaluation.

Dealing with the threats to internal validity of designs incorporating within-subjects measurements

A design incorporating within-subject measurements is exposed to several threats to its internal validity (e.g., practice, sensitivity and carry-over; Greenwald, 1976). A number of features of this study were used to try to minimize these effects:

- Several characteristics of the research were used in order to reduce the possibility that the subjects would remember their previous evaluations: (a) a 14-day interval was chosen between the two measurements; (b) each subject was asked to evaluate three teachers in each measurement; and (c) a 15-point scale was preferred over the traditional 5-point Likert scale.
- In order to deal with the possible threat of practice, two parallel groups were assigned a reversed order: one experimental group evaluated teaching using paper-and-pencil in the first measurement and online evaluation in the second, whereas the other experimental group started with the online evaluation and then did the paper-and-pencil evaluation.
- In order to deal with the possible threat of sensitivity, the research was presented as dealing with teaching evaluation and the subjects were not told that they were about to evaluate teaching again within two weeks.

Results and discussion

Of the 198 students who evaluated teaching in measurement one, 128 evaluated teaching also in measurement two (75 in the experimental group and 53 in the control

group). Most of the missing 70 students were not present in class during measurement two because of various reasons, mainly due to the different lessons in which the two measurements were taken: not all the students were registered to all of the classes. As recalled, the paper-and-pencil teaching evaluations were conducted in a regular class while online evaluations were conducted during a computer laboratory lesson.

The unit of analysis is an evaluation of teaching in a course submitted by a student. There were 338 units of analysis: 194 in the experimental group and 144 in the control group.[1]

Comparing the questions' means for the experimental and control groups

The means of the seven questions were compared in measurements one and two for the experimental and control groups. Table 1 presents the seven questions' means for measurement one and two for the experimental and control groups.

As presented by Table 1 the patterns of the questions' means are similar for both the experimental and the control group: in both groups for six out of the seven questions the means on measurement two were higher that the respective means of the first measurement. Note that the raw means of the questions in each group are not expected to be similar because the two groups were asked about different teachers.

In order to examine the similarity between the questions means in the two groups, multiple regression analysis predicted each question's mean of the second measurement from three independent variables: the question's mean on the first measurement, the group variable and the interaction between the first measurement and the group variable. The results revealed that only the first measurement was a significant predictor (beta=.974; $p<.01$) while the group variable and the interaction variable had no partial contribution (beta coefficients of .19 and –.08, respectively, both $p>.10$).

That is, no differences were found in the two measurements means due to the different mode of administrating the teaching evaluation. Such similar patterns of the questions means are expected in lieu of previous findings of between-subjects design research in the literature (Ha & Marsh, 1998; Layne *et al.,* 1999; Thorpe, 2002).

Table 1. The means of the seven questions in both measurements for the experimental and control groups

	Experimental group		Control group	
	Measurement one	Measurement two	Measurement one	Measurement two
Question 1	11.7	11.9	11.3	11.5
Question 2	11.0	11.3	10.6	10.9
Question 3	12.7	12.5	12.4	12.3
Question 4	11.2	11.5	10.4	10.7
Question 5	11.0	11.4	10.4	10.9
Question 6	11.5	11.8	11.0	11.2
Question 7	10.2	10.7	10.5	10.8

Analyzing the between measurement difference in the two groups

As recalled, the fact that the different modes of administrating teaching evaluation do not affect the evaluation's means does not indicate possible differences within the subjects. In order to examine such possible differences the absolute difference within each subject between the second and first measurement was computed for each question.

Figure 1 presents the two groups' averages of the absolute difference for each of the seven questions in the questionnaire.

As illustrated by Figure 1, in six out of the seven questions the mean absolute difference was higher in the experimental group than in the control group. Three of these six differences were statistically significant ($p<.05$). That is, using different modes of administrating teaching evaluation was usually accompanied by lower stability of the evaluations.

Additional examination of the differential stability in the two groups used the correlation coefficient between the two measurements for each question. Figure 2 presents these correlation coefficients in the two groups.

As illustrated by Figure 2, for all seven questions the interval stabilities of the two measurements as measured by the Pearson correlation coefficient were higher in the control group than in the experimental group. Two out of these seven differences were statistically significant ($p<.05$).[2] Again, the results show that using different modes of administrating teaching evaluation was usually accompanied by a lower stability of the evaluations.

Examining the possible reasons for the differential stability

In order to examine the possible reasons for the higher stability in the control group relative to the experimental groups, two variables were analyzed: the internal reliability

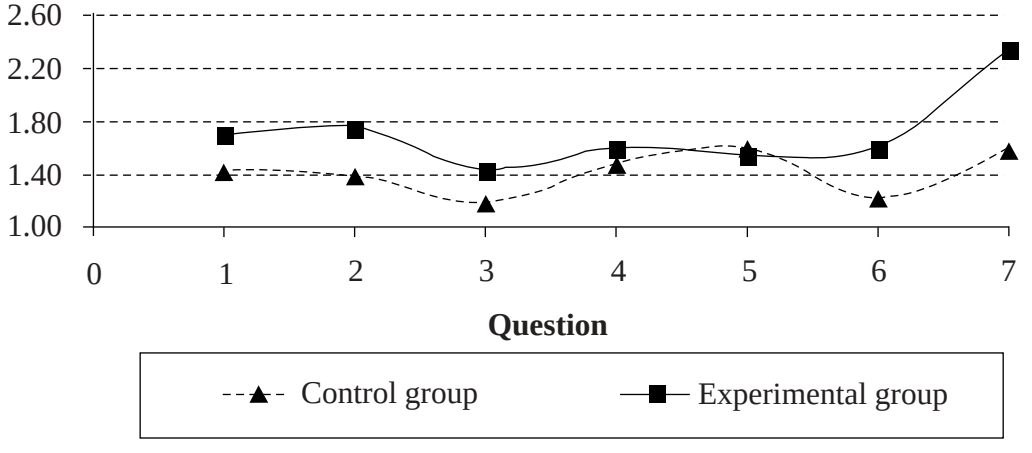

Figure 1. Mean absolute difference between the two measurements for the two groups in all seven questions

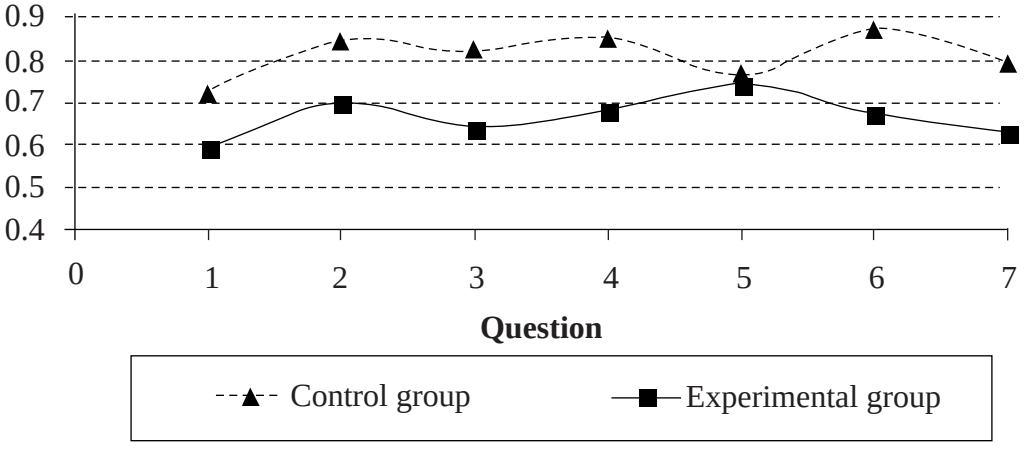

Figure 2. Correlation coefficients between the two measurements for the two groups in all seven questions

of the measurements and the variability of the answers within each subject. We expect that these variables will have differential patterns for the online method versus paper-and-pencil method. Hence, the following analysis will refer to the two subgroups of the experimental condition separately: the experimental group that evaluated teaching using paper-and-pencil on measurement one and using online evaluation on measurement two will be referred to as experimental group one (30 subjects), while experimental group two (45 subjects) evaluated teaching in a reversed order.

Internal reliability

Table 2 presents the Cronbach alpha internal reliability coefficients of the seven questions in both measurements for the two experimental groups and the control group.

Table 2 shows that the internal reliability coefficients were relatively high although the questionnaire was composed of only seven items. These values are similar to those found in the literature with respect to paper-and-pencil questionnaires. The typical Cronbach's alpha measures reported are above 0.90 and can reach 0.95 for 20-item questionnaires (Marsh, 1982; Coffey & Gibbs, 2001).

Table 2. Internal reliability coefficients for teaching evaluations on both measurements

	Measurement one	Measurement two
Experimental group one	0.92	0.89
Experimental group two	0.91	0.94
Control	0.95	0.97

Note: measurement two of experimental group one and measurement one of experimental group two used the online method; all other measurements used paper-and-pencil.

Table 2 further shows that the internal reliabilities were slightly higher for the control group on both measurements. Moreover, for both experimental groups the internal reliability coefficients were slightly higher for the paper-and-pencil method (0.92 and 0.94 for experimental group one on measurement one and for experimental group two on measurement two, respectively) relative to the online method (0.89 and 0.91, respectively).

Indeed, lower reliability can cause lower stability because reliability measures are upper boundaries for validity measures. However, the small differences found between the experimental and control groups, as well as between the two modes within the two experimental groups, imply that the differential reliabilities cannot account for all the differential stability.

Within-subjects variability of the answers to the seven questions

The between questions variability was examined for the two experimental groups and the control group. Figure 3 presents the mean of the standard deviation calculated between the seven questions for each measurement in the two experimental groups and the control group.

As presented by Figure 3, the mean of the standard deviations was higher when evaluation was conducted online: in measurement two of experimental group one (1.71) and in measurement one of experimental group two (1.76). The standard deviations for the paper-and-pencil administration in the experimental groups are lower (1.63 and 1.28) and similar values were found on both measurements in the control group (1.61 and 1.25). The difference between paper-and-pencil and online administration is statistically significant ($p<.05$). Thus, the within-subject variability of the answers is higher for subjects evaluating teaching using online versus

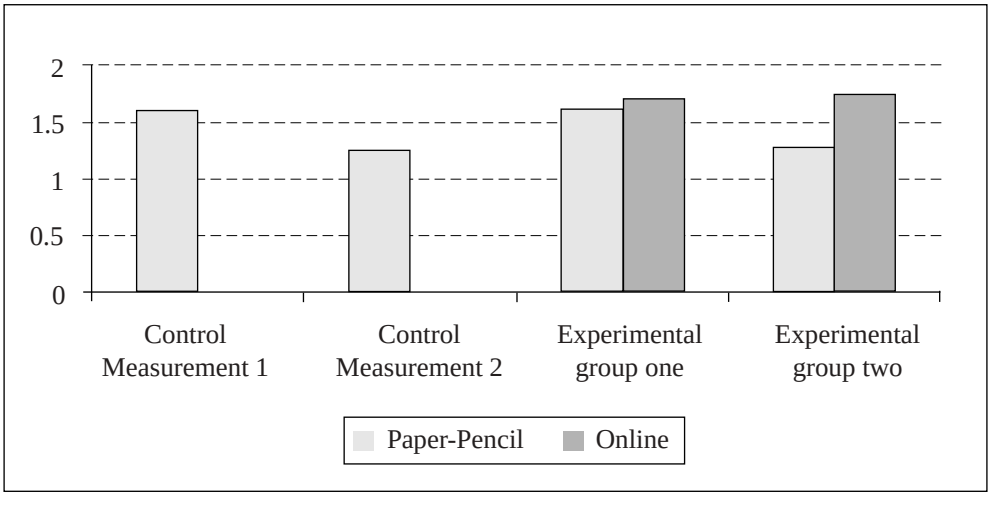

Figure 3. The means of the standard deviations between questions for each measurement in the two experimental subgroups and the control group

paper-and-pencil. This holds for the same subjects evaluating with different methods as well as for different subjects.

One possible explanation for the differential answers' variability in the two methods might be the format of the questionnaire. As recalled, the two methods differed in their visual presentation of the scale used to evaluate teaching: in the paper-and-pencil questionnaire the scale was presented horizontally on one page for all seven questions, while in the online version a vertical scale was opened for each question in a separate window that closed after the subject marked the evaluation. The visual presentation in standard paper-and-pencil forms might create a demand characteristic in which students evaluating teaching could be inhibited from giving different answers to different questions. The immediate figural feedback of the answers' profile can call for a relatively flat profile. Such figural feedback does not exist in the online evaluation form used in this study.

Conclusions

The main conclusion of this study is that the mode can matter: although the results of this study are consistent with previous findings of similar means of teaching evaluation on both paper-and-pencil and online methods, the stability of the evaluations is higher when the mode is held constant—paper-and-pencil. These results imply that the expected individual difference between the two evaluations would be higher if the two methods of evaluation were different. Our analysis suggests that a possible reason for the lower stability achieved with a different method might be the different visual presentation of the scales used in the paper-and-pencil and the online questionnaire: whereas the standard paper-and-pencil questionnaires include a horizontal scale for the questions enabling the student to obtain figural feedback on the answers' profile, the online evaluation form does not enable such a figural feedback. This explanation is further supported by the consistent higher internal reliability obtained for the paper-and-pencil method relative to the online method.

The one page standard paper-and-pencil form enables an efficient computerized scanning that has economic and logistic advantages. As shown in this study, these advantages might be accompanied by a disadvantage of artificially lowering the variability of the students' evaluations. Typical questionnaires include various questions because it is assumed that the questions tap on different aspects of teaching quality. The fact that paper-and-pencil questionnaires seem to lower the variation between these questions is problematic. Whereas solutions of this artificial bias using paper-and-pencil forms are cumbersome, an online form could easily solve this problem. Thus, an additional benefit could be added to the list of online evaluation advantages.

In this study, the student population was slightly older (average age was 33-years-old) than the typical student's age. Nevertheless, whereas the examined variable was the stability level of the various teaching evaluation methods, it seems that the age of the subjects should not affect the findings, and that the findings could be generalized

for younger student populations. This is supported by the fact that the subjects' computer and Internet literacy is not less than that of younger students, as computers and the Internet are common managerial tools.

This study used a mixed-design that compared the within-subjects differences between groups that varied in the mode of evaluation. To the best of our knowledge this is the first attempt to examine the effect of the evaluation mode by using individual differences. Additional research is required in order to try and determine whether the lower stability between the two methods found in this research results from characteristics of the online method or is due to the dissimilar visual presentation. One such research might use three experimental groups in a mixed-design: all groups would evaluate teaching twice using the same method; one experimental group would use the traditional paper-and-pencil method with the standard horizontal scale; second and third experimental groups would use online evaluation of teaching either with the standard horizontal scale or with the hidden vertical scale. If the different visual presentation were indeed the cause for the differential stability, one would expect similar stability in the first and second groups and lower stability in the third group. It should be noted, that such research should be performed using one class divided randomly into different experimental groups that evaluate the same teachers.

Notes

1. Each student was asked to evaluate three courses. However, there are fewer evaluations than 3 * 128 because not all students took all three of the courses about which they were asked.
2. Two additional differences were statistically significant at .10.

Notes on contributors

Eyal Gamliel is at the Behavioral Sciences Department at Ruppin Academic Center, Israel.

Liema Davidovitz is at the Department of Economic Management and Accounting at Ruppin Academic Center, Israel.

References

Baum, P., Chapman, K. S., Dommeyer, C. J. & Hanna, R. W. (2001) Online versus in-class student evaluations of faculty, paper presented at the *Hawaii Conference on Business*, Honolulu.

Carini, R. M., Hayek, J. C., Kuh, G. D., Kennedy, J. M. & Ouimet, J. A. (2003) College student responses to web and paper surveys: does mode matter?, *Research in Higher Education,* 44(1), 1–19.

Coffey, M. G. & Gibbs, G. (2001) The evaluation of the Student Evaluation of Educational Quality questionnaire (SEEQ) in UK higher education, research note, *Assessment & Evaluation in Higher Education,* 26(1), 89–93.

Cummings, R., Ballantyne, C. & Fowler, L. (2001) Online student feedback surveys: encouraging staff and student use, in: E. Santhanam (Ed.) *Student feedback on teaching: reflections and projections* (Australia, The University of Western Australia), 29–37.

Dommeyer, C. J., Baum, P. & Hanna, R. W. (2002) College students' attitudes toward methods of collecting teaching evaluations: in-class versus online, *Journal of Education for Business,* 78(1), 11–15.

Eckel, C. C. & Grossman, P. J. (2000) Volunteers and pseudo-volunteers: the effect of recruitment method in dictator experiments, *Experimental Economics,* 3, 107–120.

Goodman, A. & Campbell, M. (1999) *Developing appropriate administrative support for online teaching with an online unit evaluation system,* paper presented at the *International Symposium on Intelligent Multimedia and Distance Education* (Baden–Baden, Germany).

Greenwald, A. G. (1976) Within-subjects designs: to use or not to use?, *Psychological Bulletin,* 83, 314–320.

Ha, T. S., Marsh, J. & Jones, J. A. (1998) Web-based system for teaching evaluation, paper presented at the *Lingnam College Thirtieth Anniversary Conference*, Hong Kong.

Ha, T. S. & Marsh, J. (1998) Using the web for student evaluation of teaching, paper presented at the *First Conference to Promote Teaching and Learning*, Hong Kong, Polytechnic University.

Hmieleski, K. & Champagne, M. (2000) Plugging in to course evaluation, *The Technology Source,* September/October. Available online at: http//distance.wsu.edu/facultyresources/savedfromweb/pluggingin.htm (accessed 30 July 2005).

Johnson, T. (2002) Online student ratings: will students respond?, paper presented at the *Annual Meeting of the American Educational Research Association*, New Orleans, LA.

Kelly, M. & Marsh, J. (1999) *Going online with student evaluation of teaching, evaluation of the Student Experience Project* (City University of Hong Kong, Hong Kong).

Klassen, K. J. & Smith, W. (2002) From atoms to bits: using web logs to understand online instructor evaluations, *Proceedings of the Thirty-third Annual Meeting of the Decision Sciences Institute,* San Diego, CA.

Layne, B. H., DeCristoforo, J. R. & McGinty, D. (1999) Electronic versus traditional student ratings of instruction, *Research in Higher Education,* 40(2), 221–232.

Marsh, H. W. (1982) SEEQ: a reliable, valid and useful instrument for collecting students evaluations of university teaching, *British Journal of Educational Psychology,* 52(1), 77–95.

McGourty, J., Scoles, K. & Thorpe, S. W. (2002) Web-based student evaluation of instruction: promises and pitfalls, paper presented at the *Forty-second Annual Forum of the Association for Institutional Research*, Toronto, Canada.

Nulty, D. (2001) Web online feedback (WOLF): intentions and evaluation, in: E. Santhanam (Ed.) *Student feedback on teaching: reflections and projections* (Australia, The University of Western Australia), 38–41.

Sax, L. J., Gilmartin, S. K. & Bryant, A. N. (2003) Assessing response rates and non-response bias in web and paper surveys, *Research in Higher Education,* 44(4), 409–432.

Theall, M. (2000) Electronic course evaluation is not necessarily the solution, *The Technology Source,* November/December. Available online at: http://ts.mivu.org/default.asp?show=article&id=823 (accessed 21 August 2003).

Thorpe, S. W. (2002) Online student evaluation of instruction: an investigation of non-response bias, paper presented at the *Forty-second Annual Forum of the Association for Institutional Research*, Toronto, Canada.

Appendix 1. The questionnaire

Following are seven closed questions that refer to the evaluation of the course and of the lecturer. The evaluation is on a scale of 1–15: 1 being the lowest grade and 15 being the highest grade.

Very low			Low			Fair			High			Very high		
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Your evaluations are anonymous

	Very low 1 2 3	Low 4 5 6	Fair 7 8 9	High 10 11 12	Very high 13 14 15
1. To which degree was the course material clear and understandable?					
2. To which degree did the course supply you with analysis and thinking tools?					
3. To which degree did the lecturer refer to students' questions and remarks?					
4. To which degree were the lessons taught in an interesting manner?					
5. What is your general evaluation of the course?					
6. What is your general evaluation of the lecturer's teaching?					
7. What is your evaluation of the course reading material?					

Age: _____ Demographic details: Gender: M/F