What's a comment worth? How to better understand student evaluations of teaching

Winai Wongsurawat
School of Management, Asian Institute of Technology, Klong Luang, Thailand

Received July 2009
Revised February 2010
October 2010
Accepted October 2010

Abstract

Purpose – The purpose of this paper is to critically analyze the value of the written comments section on student evaluations of teaching and develop a framework to improve the interpretability of such data.

Design/methodology/approach – The paper reviews past investigations into the reliability and interpretability of student evaluations of teaching, and then constructs a framework that can potentially improve the value of data gathered from written comments.

Findings – It is shown that including information about the congruence of the comment writer's empirical ratings with those of the average class participant may help instructors separate thoughtful comments that represent the majority sentiment from attitudes of a vocal minority or those with personal biases.

Practical implications – The proposed scheme can be implemented electronically while preserving the confidentiality of the evaluators.

Originality/value – The paper offers constructive suggestions on improving the written comments section, a component of student evaluations of teaching that has so far received little systematic appraisal.

Keywords Written communications, Students, Teaching, Critical training, Training evaluation

Paper type Conceptual paper

1. Introduction

Teaching evaluations are frequently comprised of closed-end, Likert-type items plus open-end questions in which students are allowed to compose their own comments about the course and/or the instructor. While much has been written about the reliability of the former (see Kulik (2001) and Emery *et al.* (2003) for recent reviews), research on the open-end comments has been scarce to date. The written data probably have received not as much attention because they are more qualitative, and hence lend themselves less readily to statistical analysis. The fact that evaluation forms still contain written sections nevertheless attests that they continue to be viewed as relevant by students, faculty, and administrators. Indeed, some researchers (e.g. Lewis (2001) and Svinicki (2001)) claim instructors actually prefer receiving written comments rather than numerical summaries of the scores from the Likert-type items, especially when the comments are specific and constructive.

Open-end, written comments can be viewed as part of the recent movement to incorporate more qualitative evaluations into the assessment of educational instruction. Ory (2000) and Harper and Khu (2007) argue that qualitative evaluations make the assessment of faculty more accurate and comprehensive, and substantially contribute to improvements in the quality of teaching. Other types of qualitative evaluations that have been proposed in the literature include teaching portfolios, peer teaching reviews, and

Quality Assurance in Education
Vol. 19 No. 1, 2011
pp. 67-83

0968-4883
DOI 10.1108/09684881111107762

student focus groups. See Contreras-McGavin and Kezar (2007) and Latucca and Domagal-Goldman (2007) for more discussion on these methods.

Among the few studies that have systematically investigated student written comments, the findings are mixed. Ory *et al.* (1980) reported remarkably high correlations – upwards of 0.9 – between Likert scale ratings and written comments. The authors interpreted the findings as evidence supporting the reliability of students' assessments of teaching. More recently, however, Alhija and Fresko (2009) found the correlation between Likert-type ratings and written comments to be substantially lower – between 0.2 and 0.5 – on common dimensions of instructor and course characteristics. They conclude that written comments potentially contain extra information that the numerical ratings do not capture.

Hardy (2003) reports that on the whole written comments are more likely to be positive than negative. Oliver *et al.* (2007) find that students who are high achievers, female, older, and non-foreign were more likely to supply written comments on teaching evaluations. Focusing on the medium through which evaluations are delivered, Sorenson and Reiner (2003) find evidence that students tend to write longer comments when evaluations are online. Avery *et al.* (2006), however, caution that online evaluations also risk lower overall participation rates.

The factors that may render written comments unreliable are well known in the literature. Akin to problems associated with quantitative evaluations, written comments potentially contribute to grade inflation (Krautmann and Sander, 1999), popularity contests among faculty (Kulik(2001), and an erosion of academic freedom (Haskel, 1997). More specifically related to written evaluations are the problems of sample size and sample selection. Because the fraction of students supplying written comments have been documented to be as low as 10 percent (Alhija and Fresko, 2009), problems of sample selection can become serious (e.g. only a handful of angry students choose to comment). In addition, when the sample size is small it may be impossible – at least for an outsider – to differentiate between, for example, thoughtful comments and "cheap shots" from a minority of disgruntled students. Lewis (2001) offers some promising ideas for improving the interpretive value of student comments. She proposes grouping comments according to overall satisfaction levels given in the Likert scale ratings. According to Lewis (2001), if certain criticisms or suggestions arise from both satisfied and unhappy students, then they are more likely to be unbiased.

In this paper, I offer suggestions that build on the ideas discussed in Lewis (2001). First, I argue that, when samples are small and selective, gauging the reliability of each comment can provide great dividends. Assume for the moment comments are organized according to Lewis's (2001) suggestion. If for example the satisfied students write comments that are contradictory to those supplied by the unsatisfied students, is there a way to judge which set of comments is more reliable? One critical measure of reliability is how thoughtful the student was in the filling in the evaluation. The congruence between a student's responses and the class averages on questions pertaining to objective aspects of the class, I argue, may yield insight into how informed and thoughtful the evaluator was.

Second, if say only moderately satisfied students happened to write comments and they turn out to contradict one another, is there a way to judge which opinions are more representative of the overall class attitude? In this case, I argue that comments that represent the mainstream sentiment are more likely to have been written by students

whose subjective assessments of the class on various dimensions correlate well with the average class trend. Put another way, the representativeness of a comment can be judged by how typical the pattern of the writer's Likert scale ratings are. Compared to Lewis's (2001) scheme, this approach offers a more direct and quantifiable measure of representativeness.

I will proceed to illustrate these ideas by presenting and discussing a concrete example in the following sections.

2. Under-determination: an example

Questions regarding the reliability and usefulness of teaching assessments can inevitably be traced back to the need to protect the anonymity of students. The lack of data regarding characteristics of the evaluators makes it sometimes impossible to assess whether a comment is serious and representative of the majority attitude or simply thoughtless and personally biased. Such difficulties have been identified as "under-determination" problems. Arguing that data from students' evaluations of teaching cannot be used to infer instructor performance, Sproule (2002, p. 287) cites the following definition of under-determination:

- when a given model does not provide a unique or an unequivocal explanation of a body of data;
- when two or more models provide equally plausible explanations of the same data; or
- when the meaning of any set of results is up for grabs.

Reviewing a hypothetical set of written evaluations of an instructor and the four different interpretations of the results, it is easy to be persuaded that such data represent an archetype of under-determination. The set of written evaluations being:

> The teacher always had a horse blind. Never listens to students.
>
> Great teacher. Taught with conviction. Very knowledgeable of the subject.
>
> Lectures were boring. Exams were completely off topic and did not measure students' learning.

The different interpretations are:

(1) *Different backgrounds.* The instructor's performance was satisfactory. However, student backgrounds happened to be highly varied, leading to large differences in opinions about the instructor's performance.

(2) *Poor performance.* Most students found the instructor's performance to be poor. Perhaps a few inattentive enrollees were happy with the easy grading; or only top students enjoyed the class while the rest found the pace much too fast.

(3) *Satisfactory performance.* The instructor performed adequately well, but refused to cave to pressure from a few unrepresentative, but highly vocal students. Few of the satisfied majority chose to write comments.

(4) *White noise.* The contradictory comments arose from oddballs or students with personal grudges. In neither case do the comments contain much informational content.

Specifically, any of the offered models – poor performance, adequate/satisfactory performance, or white noise – appear to offer equally plausible explanations of the observed pattern. Furthermore assuming adequate/satisfactory performance, this model does not provide a unique or an unequivocal explanation of the same data, as the negative comments could be due to a highly diverse student background or simply a few troublemakers. Finally, based on these observations the interpretation of the written comments appears to be "up for grabs".

It may be possible to better identify the underlying model using additional information. For instance, a wide range of majors and GPAs of enrolled students would lend some support to interpretation 1, while abnormally high grades might be seen as evidence supporting interpretation 2. Nevertheless the sample selection problem associated with student comments makes drawing any conclusions from these data unreliable.

3. Resolving under-determination

Prior to supplying written comments, students are customarily asked to rate various attributes of the course using a Likert scale (with ratings ranging from "strongly agree" to "strongly disagree", for example). These attributes include characteristics that are relatively concrete and objective in nature as well as characteristics that are highly abstract and subjective. Examples of relatively objective evaluation criteria are:

> The course outline provided accurate description of the course.
>
> The requirements of the course (projects, papers, exams) were adequately explained.
>
> The instructor provided timely feedback to the class.

The more subjective evaluation criteria could involve the following:

> The workload was appropriate.
>
> I learned a lot from this course.
>
> The instructor stimulated my interest in the subject.

By investigating the average correlation of each writer's ratings with the mean class ratings on the same criteria, additional insights that can contribute toward resolving the under-determination problem can be gleaned. Consider Table I as an example. The first number in each cell represents the average correlation of the writer's objective evaluations with the mean ratings of the whole class on the same objective ratings. A value of 0.7 signifies that, considering only the objective criteria, the comment writer's evaluations closely resembled the patterns exhibited by the class's average ratings. The second number represents the parallel measure for ratings on subjective criteria.

It seems reasonable to expect students' ratings on objective criteria to be positively correlated because they involve less subjective, personal judgment. For example, if by the end of the course the material covered closely resembled the courses outline, but the midterm papers were not returned until the week before the final, then it seems reasonable to expect students to rate the adequacy of the course outline higher than the timeliness of faculty feedback. Such expectations appear to be fulfilled in scenarios 1 through 3. The pattern however fails to hold in scenario 4. How might this last outcome occur?

	Scenario number 1	2	3	4
"The teacher always had a horse blind. Never listens to students"	0.7, 0.2	0.6, −0.1	0.7, 0.8	0.0, 0.2
"Great teacher. Taught with conviction. Very knowledgeable of the subject"	0.6, 0.0	0.6, 0.7	0.8, −0.2	−0.2, 0.1
"Lectures were boring. Exams were completely off topic and did not measure students' learning"	0.7, −0.1	0.8, 0.0	0.6, 0.6	−0.1, 0.0
Likely interpretation	1	3	2	4

Notes: The first number in each cell is the average correlation of the writer's objective evaluations with the mean ratings of the whole class on those questions. The second number represents a similar measure for the subjective criteria

Table I.

Sproule (2002) describes a personal experience that might embody scenario 4. In one of his classes, grades were based completely on two midterms and a final exam. For each of the midterms, exams were marked, returned to students, and discussed during the class following each exam (i.e. the earliest possible date). Despite achieving the best possible record on prompt feedback, only about half the class rated the professor a 5 out of 5 on "Providing timely feedback to class". About one quarter of the class rated the instructor a 4 out of 5 and another one quarter a 3 out of 5 on this criterion. Similar stories are reported in Emery *et al.* (2003).

Such an example suggests that some students either:

- fail to pay adequate attention in class; or
- do not put sufficient thought and effort in evaluating the instructor (e.g. evaluating everything a 3 out of 5).

In either case, the ratings of such students would most likely correlate poorly with the overall patterns of the more attentive, average student. Comments in scenario 4 therefore appear to fit best with interpretation 4 (white noise). Because these students either lack the information or adequate motivation to evaluate the course, the supplied comments can thus be safely discounted or otherwise viewed with a healthy dose of suspicion.

Focusing on scenarios 1-3, all comments exhibit high correlations with the class average on objective criteria ratings. The writers therefore appear to have paid attention in class, and thoughtfully evaluated the instructor on the objective criteria. Turning to the subjective criteria correlations, scenarios 2 and 3 seem to suggest rather clear implications. In scenario 2, the positive comment was written by a student whose subjective evaluations were inline with the average class opinion (correlation of 0.7), while the negative comments were written by a non-representative minority (correlation of -0.1 and 0.0). Scenario 2 would thus appear to fit better with interpretation 3 (satisfactory performance). For the same reasons, the correlation evidence suggests interpretation 2 (poor performance) best explains the outcomes in scenario 3.

Finally, the patterns observed in scenario 1 seem to hint that there is little consensus regarding the subjective ratings of the instructor (neither the positive nor negative

comment writers' subjective ratings correlate well with the class average assessment). Assuming the objective evaluations in which a clear consensus was achieved were positive, the best explanation for scenario 1 might therefore be that the instructor's performance was adequate, but the highly diverse background of enrollees led to large differences in the subjective evaluations of students as well as in the written comments (interpretation 1).

4. A conceptual framework

Figure 1 illustrates the conceptual framework behind the example in the previous section. Each written comment will fall in one of the depicted quadrants. Those falling in the northeast and southwest are most easily appraised, the former being thoughtful remarks that resemble the average class sentiment and the latter being noise. Interpreting comments in the northwest and southeast quadrants can be more challenging. Comments falling in the southeast quadrant imply that the writer agreed with other participants on the objective evaluation dimensions (e.g. assignments clearly explained, feedbacks given in a timely manner). If one can find other comments that show strong positive correlations on the subjective criteria, then the comments with weak subjective correlations are likely a minority view on subjective matters (e.g. appropriateness of workload or teaching effectiveness). However if none of the comments exhibit strongly positive subjective correlations, then the possibility arises that there is no consensus among the class on subjective matters, possibly due to highly diverse backgrounds of the participants.

Comments that fall in the northwest quadrant also probably need to be viewed with suspicion. Although the writer arrived at similar subjective evaluations as the average participant, the failure to achieve a reasonable agreement on objective matters calls the reliability of the opinion into doubt. Possible explanations of such outcomes may be:

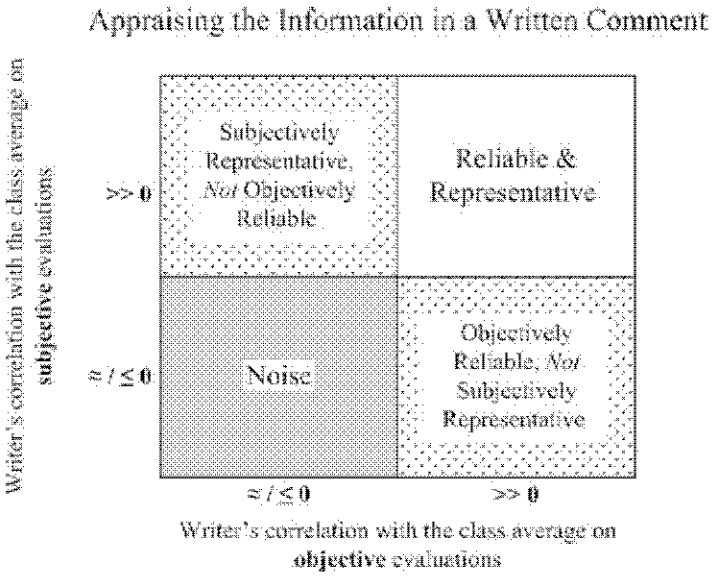

Figure 1.
A conceptual framework for judging reliability and representativeness of a comment

- the student failed to thoughtfully evaluate the objective dimensions of the class, or
- limited language abilities of participants hindered effective in-class communication.

5. Applying the idea to real data

Students enrolled in the Quantitative Methods for Management course at [name of institution withheld for confidentiality during the review process] in the August 2010 semester were requested to supply mid-semester feedback to the instructor (the author). A staff member from the Dean's office distributed and collected the paper evaluation forms at the beginning of class in the absence of the instructor. Students were instructed not to write their names on the form, and clearly informed that their participation was strictly voluntary. All 69 students enrolled chose to fill out and return the documents.

The evaluation form used (see Appendix 1, Figure A1) was an exact replicate of the institute's standard course evaluation taken at the end of each semester. The only difference was that the regular, semester-end surveys are done online. The form contains 35 Likert-type items, the scale running from 1 ("strongly disagree") to 5 ("strongly agree"). These items are subdivided into four sections (course characteristics, teaching methods and materials, instructor characteristics, and overall assessment). After each section, some space is provided for students to supply open-ended, written comments.

Studying data from the online, semester-end evaluations would admittedly be a superior option to using mid-semester, paper-based questionnaires. Because students would have observed the whole course and would perhaps be more assured of their confidentiality in an online survey, their comments might have been more accurate and thoughtful. Unfortunately, the Office of the Vice President for Academic Affairs has informed the author that past data on student evaluations of teaching have been automatically aggregated to ensure confidentiality and economize on computer storage space. It is therefore currently not possible to utilize these data to investigate the proposed framework in this paper.

Using data from the mid-semester survey, an initial review of the results suggests that students held largely positive opinions about the course. For example, in the average student ratings for the statements "You would recommend the course to another student" and "Overall, I am satisfied with the course" were 4.23 and 4.28 (out of a maximum of 5) respectively. These results are in agreement with the predominantly positive evaluation outcomes for this course over the past three years. Of the 69 respondents, only 28 supplied comments in at least one of the four sections. It is therefore natural to question whether these comments express opinions that are representative of the average student enrolled. Let us consider the following selected comments from the mid-semester survey (copied verbatim):

> The course is very practical and I am sure I can use it in future business decisions. [. . .] (The instructor) is very disciplined (Student A).
>
> Should add more topic and solve more numerical (Student B).
>
> Course topics very easy. It should be taught to make it more competitive. [. . .] Course should be more difficult. More concepts/topics should be covered in this course (Student C).

The course seems like an integral part of learning and applying management studies. Maybe more time and credit should be given to the course (Student D).

Professor should upload course materials before class start at least one day (Student E).

Everything was clearly stated on the syllabus and mentioned at the beginning of the course. [. . .] I am satisfied with the course and will take another course that builds on this topic (Student F).

Before the examinations we were told to study from what is taught in class and were told the pattern would be like last year's midterm exam. But the same did not happen. I believe specific information should not be provided if it is not going to be followed. [. . .] Writing a whole page to get your paper rechecked is probably not fair. And then having a condition that marks can be deducted seems like a threat. We are in a learning institution not in the army [. . .] The subject is really interesting and I like the prof's teaching method but I suggest him to change or be liberal during rechecking papers and providing info for exams (Student G).

The professor is really knowledgeable. Very committed to his work. I like his teaching style (Student H).

Background

- The exam from the previous year was given to students before the midterm.
- The instructor required students who wanted their exams re-graded to submit their requests in writing, explaining why they deserved higher marks. There was, however, no requirement that the written request must be at least one page long. By submitting their papers for re-grading, it was agreed that the whole paper would be reconsidered, so additional point deductions would be possible as well.

Some of the comments above provide concrete suggestions about the course logistics while others offer opinions that are sometimes positive and sometimes negative. How might one assess the reliability and representativeness of these comments? One way is to study the average Likert scale ratings on related areas. For example, if three quarters of the class "strongly agreed" or "agreed" with the Likert item "The level of difficulty in this course was appropriate", while only ten percent "disagreed" or "strongly disagreed", then it is probably reasonable to conclude that the comment about the course being too easy is unrepresentative. Nevertheless, it is not hard to imagine that in many situations such an approach would be less fruitful. For instance, conflicting comments may arise over topics that are not directly covered in the Likert items. In such situations, additional information that would help faculty or administrators judge the reliability and representativeness of a comment would potentially be very useful.

To investigate whether the ideas outlined earlier in this paper can address the challenges just described, I apply the framework summarized in Figure 1 to the mid-semester evaluation data. The Likert scale items were first categorized into those addressing objective and subjective evaluations. The objective items cover areas such as whether course logistics were clearly explained, whether course material was up-to-date, whether the instructor was accessible outside of class, and whether feedback on homework and exams was provide in a timely manner. The subjective

items covered areas such as whether the level of difficulty was appropriate, whether the instructor stimulated the student's interest, whether the student felt he or she learned a lot from the course, and whether the student would recommend this course to a friend. Admittedly, every Likert item contains some aspect of subjectivity. Yet, I believe some items (e.g. whether the level of difficulty was appropriate) are significantly more subjective than others (e.g. whether course logistics were sufficiently explained). Details about which questions were categorized as objective or subjective is provided in Appendix 1 (Figure A1). Modifying the categories would not affect the main thrust of the paper.

The correlation between the Likert scale ratings for each student and the average ratings of the class were computed separately for the subjective and objective items. Computed correlation statistics for each and every comment received is provided in Appendix 2 (Table AI). For this particular example, a correlation of between 0.5 and 1.0 was categorized as high ($>>0$) while anything between 0.2 and -1.0 was categorized as low ($\approx/\leq 0$). I have chosen to make no judgments about the reliability of comments whose Likert scale ratings show intermediate levels of congruence with the average class pattern. A technical difficulty arises for a few cases in which students' Likert scale ratings showed no variation, e.g. when a student rates every objective item as a four out of five. When such ratings occur, the correlation between the student's ratings and the average class rating involves division by zero, and is therefore undefined[1]. There are a few possible approaches to deal with this problem. One possibility would be to eliminate these problematic evaluations from the data. I have however chosen an alternative approach by assigning a correlation value of zero when there is no variation in a student's ratings. The reasoning is that a lack of variability in a student's ratings usually implies a lack of thoughtfulness in filling out the questionnaire. In extreme cases, disgruntled students may rate every aspect of the course the lowest possible score. Assigning a zero correlation to these evaluations ensures that comments from such students are not given priority.

Summary results from applying the analytical framework in Figure 1 to the selected set of student comments above are given in Figure 2. The analysis indicates that the comments about the course warranting more time and credit (written by student D), and the suggestion that the slides be posted earlier (written by student E) are likely reliable and representative. On the other hand, the comments about the course needing more material (written by student B) or being too easy (written by student C) should not be given much weight.

Also interesting is the assessment of student G's very long comment. Here, the analysis indicates that the student's objective evaluations are in-line with the class, but his or her subjective views may not be highly representative. Student G's opinions about the class being run like an army or that the instructor should be more "liberal" in grading exams may therefore be legitimately discounted.

It is important to note that the proposed analysis does not always provide a neat and simple categorization of "useful" and "useless" comments. Indeed, similar comments may show up in multiple quadrants of the appraisal matrix. In the present example, compliments about the instructor's enthusiasm as well as complaints about slides not being uploaded early enough show up multiple times in comments that my framework judges to be both representative and non-representative. Also a possibility is that two directly contradicting comments may both be judged by the correlation

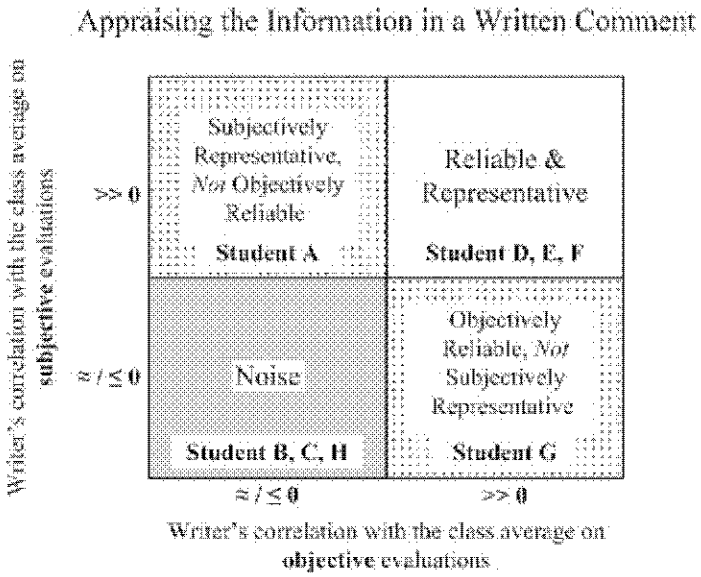

Figure 2.
A conceptual framework for judging reliability and representativeness of a comment

analysis to be equally reliable and representative. It is therefore best to view this framework as a channel to provide supplementary information for improving our understanding of students' written comments. Such analysis certainly is not a substitute for detailed understanding of what happened in class and whether issues arising from student comments warrant serious consideration for future improvement.

6. Concluding remarks

Because of the flexibility they afford, written sections on student evaluations of teaching potentially contain rich information that could be useful to instructors (for improving future performance), administrators (for evaluating teaching staff), and students (for course selection, in cases where the evaluation results are made available to students as well). However, the anonymous format of these evaluations commonly used today suffers from the under-determination problem. Specifically, a collection of usually contradictory comments lend themselves to many equally plausible, yet fundamentally opposite interpretations.

Because anonymity is a necessary requirement in evaluations in which the party being evaluated is in a position of power, I propose a method for gauging the reliability and representativeness of each comment based on the correlations between the writer's and the class-average empirical assessments of the course characteristics. Correlation can be measured along two dimensions, the objective and subject evaluation criteria. Congruence on the objective dimension ensures that the writer of the comment paid attention in class and thoughtfully filled out the evaluation form. Congruence on the subjective dimension increases the likeliness that a subjective opinion represents a majority sentiment as opposed to feelings of a vocal minority.

The proposed scheme supplies potentially useful information for judging each written comment without compromising the anonymity of the writer. Furthermore, this approach circumvents the well-known problem that only a handful of students

usually write comments for class evaluations. Because credibility is appraised based on congruence with the average class opinion (obtained from answers in the empirical, Likert scale sections), a single representative comment, for example, can be ruled more reliable than two or three quirks. Finally, the approach makes it more difficult for students to take cheap shots at instructors.

The framework was applied to mid-semester feedback data supplied by 69 students enrolled in a master-level course. Although the received comments were mostly positive and compatible (thus making the comments less enigmatic and perhaps rendering the correlation analysis less vital in this case), the framework still yielded interesting insights that mostly confirmed the instructor's prior intuition. The correlation analysis thus shows promise in improving the interpretability of written comments. Nonetheless the framework should be viewed as a channel to provide supplementary information on students' written comments, and should not be viewed a substitute for detailed understanding of what happened during class.

Caveats of the proposed approach may include the following. First, attaching a pair of correlations to each comment may make it easier for instructors to disregard valid, minority opinions. Second, if students know that their comments will be evaluated based on how well their Likert scale ratings correlate with the class average, they may try to answer the Likert scale questions to achieve a high correlation with the class average instead of putting down their honest assessments. Finally, correlation patterns in some cases may be confusing and hard to interpret thus leading to only negligible improvements from the status quo.

Note

1. The correlation between two variables is defined as the covariance divided by the product of the standard deviations of the two variables.

References

Alhija, F.N. and Fresko, B. (2009), "Student evaluation of instruction: what can be learned from students' written comments", *Studies in Educational Evaluation*, Vol. 35, pp. 37-44.

Avery, R.J., Bryan, W.K., Mathios, A., Kang, H. and Bell, D. (2006), "Electronic course evaluations: does an online delivery system influence student evaluations?", *Journal of Economic Education*, Vol. 37 No. 1, pp. 21-37.

Contreras-McGavin, M. and Kezar, A.J. (2007), "Using qualitative methods to assess student learning in higher education", *New Directions for Institutional Research*, No. 136, pp. 69-79.

Emery, C.R., Kramer, T.R. and Tian, R.G. (2003), "Return to academic standards: a critique of student evaluations of teaching effectiveness?", *Quality Assurance in Education*, Vol. 11 No. 1, pp. 37-46.

Hardy, N. (2003), "Online ratings: fact and fiction", *New Directions for Teaching and Learning*, Vol. 96, pp. 49-59.

Harper, S.R. and Khu, G.D. (2007), "Myths and misconceptions about using qualitative methods in assessment", *New Directions for Institutional Research*, No. 136, pp. 5-14.

Haskel, R.E. (1997), "Academic freedom, tenure, and student evaluation of faculty: galloping polls in the 21st century", *Education Policy Analysis Archives*, Vol. 5 No. 6, pp. 36-9.

Krautmann, A.C. and Sander, W. (1999), "Grades and student evaluations of teachers", *Economics of Education Review*, Vol. 18, pp. 59-63.

Kulik, J.A. (2001), "Student ratings: validity, utility, and controversy", *New Directions for Institutional Research*, No. 109, pp. 9-25.

Latucca, L.R. and Domagal-Goldman, J.M. (2007), "Using qualitative methods to assess teaching effectiveness", *New Directions for Institutional Research*, No. 136, pp. 81-93.

Lewis, K.G. (2001), "Making sense of student written comments", *New Directions for Teaching and Learning*, Vol. 96, pp. 49-59.

Oliver, B., Tucker, B. and Pegden, J.-A. (2007), "An investigation into student comment behaviors; who comments, what do they say, and do anonymous students behave badly?", paper presented at the Australian Universities Quality Forum, Hobart, 11-13 July.

Ory, J.C. (2000), "Teaching evaluation: past, present, and future", *New Directions for Teaching and Learning*, Vol. 83, pp. 13-18.

Ory, J.C., Braskamp, L.A. and Pieper, D.M. (1980), "Congruency of student evaluative information collected by three methods", *Journal of Educational Psychology*, Vol. 72 No. 2, pp. 181-5.

Sorenson, D.L. and Reiner, C. (2003), "Charting the uncharted seas of online student ratings of instruction", *New Directions for Teaching and Learning*, Vol. 96, pp. 1-24.

Sproule, R. (2002), "The underdetermination of instructor performance by data from the student evaluation of teaching", *Economics of Education Review*, Vol. 21, pp. 287-94.

Svinicki, M.D. (2001), "Encouraging your students to give feedback", *New Directions for Teaching and Learning*, Vol. 87, pp. 17-24.

Appendix 1. Anonymous feedback – Quantitative Methods for Management (August 2010)

Course Characteristics	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1. The course objectives were clearly explained.	0%	0%	10%	51%	39%
2. The course outline provided accurate description of the course.	0%	1%	7%	59%	32%
3. Course topics were dealt in sufficient depth.	0%	3%	16%	64%	17%
4. The course fulfilled my expectations.	0%	1%	23%	52%	23%
5. Course requirements (projects, exams) were adequately explained.	0%	3%	16%	55%	26%
6. The level of difficulty in this course was appropriate.	6%	3%	14%	57%	19%
7. The workload was appropriate.	0%	3%	9%	59%	29%
8. The grading policy was clearly explained.	0%	0%	4%	48%	48%
9. Assessment methods were appropriate and effective.	0%	0%	14%	48%	38%

Comments on Course Characteristics:

Course Delivery/Teaching Methods /Resource Materials	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
10. The teaching methods were effective.	0%	0%	3%	58%	38%
11. The course materials were up-to-date, well prepared and useful.	0%	1%	12%	52%	35%
12. The course materials were presented in an organized manner.	0%	0%	6%	43%	49%
13. The use of information technology teaching resources helped the delivery of course material.	0%	1%	10%	45%	42%
14. The assignments were relevant and useful.	0%	1%	7%	51%	39%
15. The course materials were adequate for learning the subject matter.	0%	4%	12%	57%	28%

Comments on Course Delivery/Teaching Methods/Resource Materials:

(Continued)

Figure A1.

Instructor Characteristics	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
16. Was effective in achieving the course objectives.	0%	0%	3%	75%	22%
17. Stimulated my interest in the subject.	0%	4%	14%	51%	30%
18. English communication skills are satisfactory.	0%	0%	1%	38%	61%
19. Was responsive toward student needs.	0%	1%	9%	49%	41%
20. Treated student with due respect.	0%	0%	4%	33%	62%
21. Was accessible to students outside the class.	0%	0%	23%	61%	16%
22. Provided timely feedback to the class.	0%	1%	7%	58%	32%
23. Overall performance of instructor was satisfactory.	0%	0%	0%	39%	61%
24. Enthusiastic toward teaching.	0%	0%	1%	43%	55%
25. Sufficient knowledge of subject area	0%	1%	1%	32%	65%
26. I would recommend this instructor to other students.	0%	0%	1%	35%	61%

Comments on Instructor Characteristics:

Overall Assessment	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
33. I learned a lot from this course.	0%	3%	12%	57%	29%
34. You would recommend the course to other students.	0%	1%	14%	43%	41%
35. Overall, I am satisfied with the course.	0%	3%	9%	46%	42%

Overall Assessment/Areas for Improvement:

Figure A1.

Note: Questions 1, 2, 5, 8, 11, 12, 13, 18, 21, and 22 were categorized as 'objective' evaluations. The remaining questions were categorized as 'subjective'

Appendix 2

Table AI lists all comments received from the mid-semester feedback survey for the *Quantitative Methods for Management* course in August 2010, along with correlations with the average class Likert scale ratings on objective and subjective questions.

Comments	Corr. with average ratings on objective questions	Corr. with average ratings on subjective questions
Very interesting course indeed! However, one could argue the relevance/applications of it in real-world applications … No complaints. Good job! … Dedicated, knowledgeable with a high level of professionalism …	0.75	0.47
Very good course … Good instructor …	0.23	0.44
This course will help to become efficient manager … Teaching method is unique. Fulfilled student's all queries using real life example … He is strict in principle. This characteristic enchanted me … The teacher may find out the relatively poor student and pick them up in the interactive class	0.36	0.62
Things might be tough at the beginning but at the end of the day you would feel value added to yourself … Since the semester is modular, it is diff to explain in depth but is good … Good professor … I am satisfied	0.61	0.14
The course would need more exercise in class and by student also … Good, excellent …	0.61	0.50
The course seems like an integral part of learning and applying management studies. Maybe more time and credit should be given to the course … Course delivery and teaching methods were good enough to understand the course …	0.79	0.62
The course is very practical and I am sure I can use it in future business decision making … Very good. Like the way he teaches … Very disciplined … Very useful	0.20	0.67
Quant analysis helped me understand the importance of decision making applications in the organization … Teaching method were sufficiently delivered and captured. Access of the server were easy … Clear in his instructions …	0.15	0.40
It is too difficult for me. I try my best but can't understand the subject. There should be some arrangement for students like me to pass the course. I mean apart from exams, something like a 20 page assignment or so … Course delivery was ok. Teaching method was nice. Resource material were very difficult … He is a nice human being with very clear mission and vision. But sometimes he is too rigid and does not show any flexibility … More flexibility: all the students are not same. Different human beings has different needs. Their capacity of learning is also not same. Keeping this in mind, there should be some support system for students who are performing poor in exams. It is not fair to evaluate an engineer, who has learnt these things earlier with a student with performing arts/humanities/environmental knowledge. So we need some system to compensate that. Maybe in this way if some one gets below C + then only they are eligible to submit special assignment because they are weak	0.28	0.72

(*continued*)

Table AI.

Comments	Corr. with average ratings on objective questions	Corr. with average ratings on subjective questions
It is a good and useful subject, but putting it in the module system just rushes things … He is a very good instructor. You may not think so in the beginning but later you would agree … Module system makes this course a hassle that students struggle with.	0.31	0.67
Instructor is good for this subject … Should add more topic and solve more numerical … The good course that helps in making us know how to take decision in tough time with options in hand … Teaching method is good enough along with resource material	0.09	−0.11
In spite not having quantitative background, I was able to understand … The material was sufficient and clear cut to understand the course. I would recommend this instructor to other students … I like the teaching style and I am satisfied with the material	0.48	0.62
His teaching is simply excellent … Course should be more difficult. More concepts/topics should be covered in this course … Course looks very easy. It should be taught to make it more competitive … Should be more difficult.	0.00	0.15
Had to digest the concepts within a small period of time … OK … Good! … Good! … Good! … Very good!	−0.39 0.65	0.51 0.40
Exam questions were a bit harder than the thing he asked us to do in class and practice exam …	0.38	0.26
Everything was clearly stated on the syllabus and mentioned at the beginning of the course … The instructor makes the course very fun and interesting. He has an effective way at teaching the course … I am satisfied with the course and will take another course that builds on this topic/course	0.76	0.65
Difficult but interested … Good … Good … Good Course is challenging and essential … Good and sufficient course material … Good, enthusiastic … Should a lot some time for tutorial sessions as 30 hours are insufficient for the whole course with problems discussion	0.67 0.61	0.35 0.44
Course content is a bit less. But the exam question was not as told in class … Teaching methodology by professor is quite satisfactory to me. Thank you … Instructor is very good to me … Improvement is always necessary	0.50	0.47
Before the examination we were told to study from what is taught in class and were told the pattern would be like last year's midterm exam. But the same did not happen. I believe specific information should not be provided if it is not going to be followed … Very clear and concise teaching style … Writing a whole page to get your paper rechecked is probably not fair. And then having a condition that marks can be deducted seems like a threat. We are in a learning institution not in the army … The subject is really interesting and I like the prof's teaching method but I suggest him to change or be liberal during rechecking papers and providing info for the exams	0.65	0.00

Table AI.

(*continued*)

Comments	Corr. with average ratings on objective questions	Corr. with average ratings on subjective questions
… The way of teaching is step-by-step and makes students clear. It is excellent teaching …	0.00	0.50
… The professor is really knowledgeable. Very committed to his work. I like his teaching style …	0.00	−0.31
… Professor should upload course materials before class start at least one day …	0.76	0.52
… Please provide slide before class begins (more than 1 hour). Students want to get it before studying …	−0.02	0.08
… I would like to get ppt before the class start at least for a day …	0.62	0.42
… Very good instructor …	0.40	0.59
… The exams were not as explained in the class. He flipped the coin completely …	0.00	0.26

Table AI.

Corresponding author
Winai Wongsurawat can be contacted at: winai@ait.ac.th