Quantitative assessment of course evaluations.

Tamara Sliusarenko

Kongens Lyngby 2013
PhD-2013-318

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Matematiktorvet, building 303B, DK-2800 Kongens Lyngby, Denmark
Phone +45 4525 3031 , Fax +45 4588 1399
reception@compute.dtu.dk
http://www.compute.dtu.dk/ PhD-2013-318

Summary (English)

Student evaluation of teaching has been used in educational institutions around the world as a means of providing feedback on the quality of teaching. Nowadays, it is one of the most widespread tools used to inform teachers and administration about the instruction given in an institution.

The goal of the thesis is to develop efficient tools to analyze the data from student evaluations of teaching and courses at the Technical University of Denmark.

The thesis explores both classical and modern methods of multivariate statistical data analysis to address different issues of student evaluation of teaching (SET). In particular, the thesis includes results on the investigation of the association between the student evaluations of the course and the student evaluations of the teachers, the investigation of the effects of the mid-term evaluation on the end-of-term evaluations and the investigation of the student non-response on SETs. In order to utilize information from open-ended qualitative student answers, text-mining methods were applied in order to extract points of students praise and complaints.

The methods proposed contribute to the knowledge about student evaluation at the Technical University of Denmark. The results provided some new information that will help teachers and university managers to better understand results of course evaluations.

Mid-term course evaluation was found to be able to capture both types of course issues: issues that can be addressed during the semester and also issues that can only be addressed at the next semester. Therefore, it seems to be preferable

to conduct general mid-term evaluations instead of end-of-term evaluation, so the current course students can benefit. Additionally, it might be beneficial to conduct a short end-of-term evaluation with very limited number of questions that focus on general course issues after the final exams in order to obtain student feedback on the entire teaching and learning process, including the alignment of assessment of students' learning with course objectives and teaching activities.

Student-specific and course-specific characteristics was found to be related with whether students participate in SETs and with how students evaluate courses and teachers. The DTU administrations should be aware that high achievers are more likely to participate in course evaluation survey and are more likely to give higher scores to courses. Students diversity on the course should be taken into account while making comparisons of evaluation results between courses.

In the student written feedback was found be able to provide additional knowledge of student point of satisfaction or dissatisfaction. However, in order to build an automated tool that can help to extract patterns from student comments higher quality of the collected data is needed.

Summary (Danish)

Studerendes evaluering af undervisningen har været anvendt i uddannelsesinstitutioner rundt om i verden som et hjælpemiddel til at give feedback på kvaliteten af undervisningen. I dag er det et af de mest udbredte værktøjer, der anvendes til at informere lærere og administration om instruktion i en institution.

Målet med afhandlingen er at udvikle effektive værktøjer til at analysere data fra de studerendes evalueringer af undervisningen og kurserne på Danmarks Tekniske Universitet (DTU).

Afhandlingen udforsker både klassiske og moderne multivariat statistisk metoder til at løse forskellige spørgsmål om de studerendes evaluering af undervisningen. Afhandlingen indeholder resultater fra en undersøgelse af sammenhængen mellem de studerendes evalueringer af et kurs og de studerende evalueringer af læreren, en undersøgelse af virkningerne af midtvejsevalueringer målt ved slutevalueringer og en undersøgelse af manglende svar på evalueringen. For at udnytte informationen i besvarelse på åbne kvalitative spørgsmål blev text mining metoder anvendt for at finde relevante ros og klager fra eleverne.

De foreslåede metoder bidrager til viden om de studerendes evalueringer på Danmarks Tekniske Universitet. Resultaterne giver nogle nye oplysninger, som vil hjælpe lærere og universitets ledere til bedre at forstå resultaterne af kursusevalueringer.

Midtvejsevaluering var i stand til at fange begge typer kursus emner: spørgsmål, der kan løses i løbet af semestret, og også spørgsmål, der kun kan løses det efterfølgende semester. Derudover kan det være en fordel at foretage en kort

slutningenevalueringen med meget begrænset antal af spørgsmål, som fokuserer på de generelle kursus problemer efter den endelige eksamen for at opnå studerendes tilbagemeldinger på hele undervisnings-og læringsprocessen, herunder tilpasning af vurdering af elevernes læring med kursets mål og undervisningsaktiviteter.

Studenter- specifikke og kursus- specifikke egenskaber var forbundet med, om de studerende deltager i evalueringerne og med, hvordan eleverne vurderer kurser og lærere. DTU forvaltningen bør være opmærksom på, at gode studerende (med høje karakterer) er mere tilbøjelige til at deltage i kursusevalueringer, og er mere tilbøjelige til at give en højere score på kursere. De studerendes mangfoldighed på kurset bør tages i betragtning, når sammenligning af evalueringsresultaterne mellem kurser skal foretages.

De studerendes skriftlige feedback gav yderligere viden om de studerendes tilfredshed eller utilfredshed. Men for at opbygge et automatiseret værktøj, der kan hjælpe med at udtrække mønstre fra studenter kommentarer vil en højere kvalitet af de indsamlede data nødvendig.

Preface

This thesis was prepared at the Section of Statistics and Data Analysis of department of Applied Mathematics and Computer Science at the Technical University of Denmark (DTU), in partial in partial fulfilment of the requirements for acquiring the Ph.D. degree in Applied Mathematics.

The work herein represents selected parts of the research work carried out in the Ph.D. time period. The project deals with different aspects of student evaluations of courses and teaching quality th DTU. The thesis consists of five research papers, one report and an introductory part containing an overview of the thesis and background information.

The project was supervised by Professor Bjarne Kjær Ersbøll (DTU) and Associate Professor Line H. Clemmensen (DTU). Part of the research was conducted at the Aalto University, Espoo, Finland, under the supervision of Professor Timo Honkela.

Lyngby, 28-October-2013

Tamara Sliusarenko

Papers included in Thesis

Chapter 7

Tamara Sliusarenko and Bjarne Kjær Ersbøll. *Canonical Correlation Analysis of course and teacher evaluations.* Paper presented at the 2nd International Conference on Computer Supported Education, 7-10 April, 2010, Valencia, Spain. (peer reviewed)

Chapter 8

Tamara Sliusarenko and Line H. Clemmensen. *How do student evaluations of courses and of instructors relate?* Paper submitted to Journal of Educational and Behavioral Statistics.

Chapter 10

Tamara Sliusarenko, Line H. Clemmensen and Bjarne Kjær Ersbøll. *Text mining in students' course evaluations: Relationships between open-ended comments and quantitative scores.* Paper presented at the 5th International Conference on Computer Supported Education, 5-8 May, 2013, Aachen, Germany. (peer reviewed)

Chapter 11

Line H. Clemmensen, Tamara Sliusarenko, Birgitte Lund Christiansen and Bjarne Kjær Ersbøll. *Effects of mid-term student evaluations of teaching as measured by end-of-term evaluations: An empirical study of course evaluations.* Paper presented at the 5th International Conference on Computer Supported Education, 5-8 May, 2013, Aachen, Germany. (peer reviewed)

Chapter 12

Tamara Sliusarenko, Line H. Clemmensen, Rune Haubo Christensen and Bjarne Kjær Ersbøll. *Respondent Bias in Course Evaluations and its Consequences* Paper submitted to Research in Higher Education.

Other works included in Thesis

Chapter 9

Tamara Sliusarenko. *Clustering the students comments.* Report for Summer school Matrix Methods for Data Mining and Pattern Recognition, August 23 - 27, 2010, DTU

Other works not included in Thesis

Bjarne Kjær Ersbøll, Tamara Sliusarenko and Line H. Clemmensen. *Does an Association Between Student evaluations of Related CDIO Courses Exist?* Paper presented at the 7nd International CDIO Conference, 20-23 June 2011, Technical University of Denmark, Copenhagen, Denmark.

Glossary

There are many terms used in the literature to describe student evaluations of course and teaching quality. Among the most common are "student evaluations", "course evaluations", "student evaluations of teaching (SETs)", "student ratings of instruction".

Each of these phrases has a slightly different connotation, depending on whether the author emphasizes the student, courses, ratings, or evaluation.

Wright 2006 suggested that the most appropriate term for end-of-course summative evaluations used primarily for personnel decisions (and not for teaching development) is "student ratings of instruction" because this most accurately reflects how the instrument is used.

Throughout this thesis, several of the above mentioned terms are used interchangeably.

Acknowledgements

I would like to acknowledge several people who put their time and effort into this project. First, I would like to thank all my colleagues from the Statistics and Data Analysis section at DTU Compute for providing an inspiring working environment. Christina Horn Nexø for guidance in the administrative part of the Ph.D program.

A special thanks goes to my supervisor Bjarne Kjær Ersbøll for encouraging guidance and supervision throughout the project to and to my co-supervisor Line Harder Clemmensen, for always being supportive with relevant feedback. I would also like to the acknowledge the Dean of Undergraduate Studies and Student Affairs of DTU, Martin Vigild, for supporting the project and the Senior Vice President and Dean of Graduate Studies and International Affairs Martin P. Bendsøe for providing the funding.

Many thanks to all the teachers and students who participated in the study, and a special thanks to Per Bruun Brockhoff and Mads Peter Sørensen for sharing their ideas and experience about the courses they are teaching. I would also like to thank the LearningLab DTU, and especially Birgitte Lund Christiansen, for assistance in carrying out the study and to the CampusNet support team especially Anders Mørup Hermansen and Christian Westrup Jensen, project manager from Study Programmes and Student Affairs, for always being helpful with various data issues.

Many thanks to Karl Sjöstrand and Rune Haubo Christensen for discussions and questions about data analysis and to Timo Honkela and Mari-Sanna Paukkeri from the Department of Informatics and Mathematical Modeling, Aalto Univer-

sity for helping me to understand and apply the text-mining methods.

I would also like to thank my examiners, Professor Knut Nonradsen from the Thechnical University of Denmark, Senior Lecturer Tom W. Adavi from Chalmers University of Technology, and Professor Lauri T. E. Malmi from Aalto University, who provided encouraging and constructive feedback. It is no easy task, reviewing a thesis, and I am grateful for their thoughtful and detailed comments.

Thanks to my husband, my daughter and my mother for their love, help and support, especially during the thesis writing period.

Contents

Part I

Introduction

Chapter 1

Introduction

This thesis analyses various aspects of student evaluations of courses and teaching quality based on the evaluations at The Technical University of Denmark (DTU). It also considers how the evaluation results can be used for future improvement of the course evaluation system and also for various university administrative decisions at DTU.

1.1 Effective teaching

One of the goals of universities and other education institutions is to provide the students with the necessary tools and instruction to meet their full learning potential. This is why so much emphasis is placed on ensuring effective teaching year after year. The most fundamental questions for education institution administrators are:

- What makes a teacher effective?
- How to measure effective teaching?

Administration, teachers and students are all interested in effective teaching. Students would like to get better knowledge. From administration's point of

view a successful school means higher national and international ranking and more students, who generate income. From a teacher's point, of view a successful school has better resources and probably higher salaries.

However, there can be no single definition of the effective teacher. The definition of effective teaching is different from the points of view of teachers, students, and administration. Moreover, the perception of effective teaching varies with the age of the student population, the background of the students, and the subject matter etc.

Leslie (2012) tried to determine if there is a difference in how "effective teaching" is defined by asking those concerned (teachers, students, and administrators) what the term means to them. When asked to select the top four out of 30 characteristics of effective teaching in order of importance, students, teachers, and administrators agreed on the same three — cultivate thinking skills, stimulate interest in the subject, and motivate students to learn — but not in the same order.

Feldman (1988) conducted a meta-analysis of 31 studies in which teachers and students identified characteristics they associated with good teaching and effective instruction. Students and faculty were generally similar, though not identical, in their views. Students emphasized the importance of teachers being interesting, having good elocutionary skills, being available, and helpful. Faculty placed more importance on being intellectually challenging, motivating students, setting high standards, and encouraging self-initiated learning.

Good teaching can't happen without student learning and without good course administration. Therefore effective teaching include teachers, the students, the curriculum, teaching methods, assessment procedures, the climate created through interactions, and the institutional climate. Particularly important are curriculum, teaching methods, and assessment procedures. When there is alignment between what we want, how we teach and how we assess, teaching is likely to be more effective (Biggs, 2003).

For administrators, teacher and students it is important to be sure that teaching is effective, therefore they try to find an appropriate measure of such a multidimensional process that combines different aspects of teaching. Berk (2005) in his study discusses 12 possible strategies to measure teaching effectiveness:

Student evaluation of teaching (SET) or student ratings, is a paper or electronic questionnaire, which requires quantitative and/or qualitative answers to a series of questions.

Peer ratings, that is peer observation of in-class teaching performance and/or

peer review of the written material used in a course.

Instructor self-evaluation is when faculty are asked to evaluate their own teaching.

Videos of the lectures provide authentic picture of how teachers really teach. Videos can be a part of instructors' self-evaluation.

Student interviews are group interviews with students that can provide more accurate, trustworthy, useful and comprehensive evidence.

Alumni ratings provide an information of what students actually remember about their instructors' teaching and course experiences.

Employer ratings helps to collect the information of what students really learn from their study programs. After some time has passed, an assessment of the graduate's on-the-job performance can provide feedback on overall teaching quality and program design.

Administrator ratings. Associate deans, program directors, or department heads can evaluate faculty for annual merit review according to criteria for teaching, scholarship, service, and/or practice.

Teaching scholarship combines presentations, publications and research in teaching and learning on innovative teaching techniques and related issues.

Teaching awards. A nominee must go through an evaluation by a panel of judges.

Learning outcome measures. It can be tests, quizzes, exams and other graded course activities that are used to measure student performance and/or students rating of their own learning.

Teaching portfolios is "a coherent set of materials, including work samples and reflective commentary on them, compiled by a faculty member to represent his or her teaching practice as related to student learning and development" (Cerbin and Hutchings, 1993).

Berk (2005) concludes that SETs is a necessary, but not a sufficient, source of evidence of teaching effectiveness for both formative and summative decisions. Peer ratings of teaching performance and materials is the most complementary source of evidence to student ratings.

Historically, student ratings have dominated as the primary measure of teaching effectiveness for the past 30 years (Seldin, 1999). Over the past decade there has been a trend toward augmenting those SET results with other sources of measuring teaching performance. Such sources can help to broaden and deepen

the evidence base used to evaluate courses and assess the quality of teaching (Arreola, 2000; Braskamp and Ory, 1994; Knapper and Cranton, 2001).

1.2 Student Evaluation of Teaching

Measures of Effective Teaching (MET) project (2012) stated that only recently have many policymakers and practitioners come to recognize that when asked the right questions, in the right ways students can be an important source of information on the quality of teaching and the learning environment in individual classrooms.

Student evaluations of teaching (SET) are collected in many education institutions all over the world. The purpose of the student ratings is mainly to provide:

- A qualitative and/or quantitative feedback to faculty and instructors for improving teaching for future students. Teachers can review how their students interpret their teaching methods, thereby improving their instruction.
- A measure of teaching effectiveness, that provides information for personnel decisions, like promotions, salary rise, tenure, reappointment, and for making formative recommendations (e.g., identify areas where a faculty member needs to improve);
- Information for potential students during the selection of courses and instructors;
- A component for national and international quality assurance exercises, in order to monitor the quality of teaching and learning;
- An outcome or a process description for research on teaching (e.g., studies designed to improve teaching effectiveness and student outcomes, effects associated with different styles of teaching, perspectives of former students).

The first two roles of SETs ae sometimes called formative and summative roles. Centra (1993) indicated, that SETs serve a formative purpose only when following conditions are satisfied:

- teachers must learn something new from evaluation results.

- teachers must value the new information.
- teachers must understand how to make improvements.
- teachers must be motivated to make the improvements

Typically, SETs are usually combined with peer evaluations, supervisor evaluations, students performance in order to create an overall picture of teaching performance.

Teaching is a complex activity with many interrelated components, like organization, teaching style, presentation skills, clarity, interaction with students, enthusiasm, ability to motivate students, feedback to students, etc. Therefore, university administrations try to construct the student evaluation questionnaire to reflect this multidimensionality of teaching (Abrami and d'Apollonia, 1991; Cashin and Downey, 1992; Feldman, 1997). According to Marsh and Roche (1997), the strongest support for the multidimensionality of SETs is based on the nine factors: Learning/Value, Instructor Enthusiasm, Organization/Clarity, Group Interaction, Individual Report, Breadth of Coverage, Examinations/Grading, Assignments/Readings, and Workload/Difficulty.

Student evaluations of teaching have been widely criticized, especially by teachers, for not being accurate measures of teaching effectiveness (Emery et al., 2003). Some of the teachers argues that SETs are biased in favour of certain teachers' personalities, looks, disabilities, gender and ethnicity and that factors other than effective teaching are more predictive of favourable SETs.

The quantity of research is indicative of the importance of SETs in higher education. A lot of issues of course evaluation have been discussed like validity and reliability of student ratings, faculty, administration and student perceptions of course evaluation and various kinds of biases in SETs. Many researchers have stated that student rating is the most valid and practical source of data on teaching and course effectiveness (McKeachie, 1997). It is now standard for universities to introduce student evaluations of courses and teaching.

1.3 Motivation

Course evaluation questionnaires usually consist of two parts. A quantitative part, where students can give numerical rating to some aspect of teaching, and a qualitative part, where students can write their feedback in words.

The numerical section of the survey is important for a number of reasons:

First, it is easy to complete and requires little effort from the student. It is observed that the shorter and easier a survey is, the higher the response rate.

Second, the numerical part ensures that all students are asked exactly the same questions, which then makes it easier to compare student's responses, within the class as well as between classes, instructors, programs, departments or perhaps even universities.

Third, the student ratings can be subjected to a variety of statistical calculations and modelling. However, there are limitations of the numerical part of the survey. While it makes comparisons easier, and provides excellent quantitative data, there is a limit to the detail that can be identified from the data.

The primary purpose of student comments is to give an individual feedback to teachers, or for use in one-to-one evaluations with administrators for personnel decisions.

Students at The Technical University of Denmark (DTU) regularly evaluate courses by filling in the so-called: "final-evaluation" web-forms on the Intranet CampusNet . The on-line course evaluation is performed a week before the final week of the course and consists of three forms:

Form A contains quantitative questions about the course, like course workload, content, teaching materials, etc.

Form B contains quantitative questions about each individual teacher, like teacher's communication, motivation, feedback, etc.

Form C gives students the possibility to provide qualitative feedback on 3 questions:

1. What went well?
2. What did not go so well?
3. Suggestions for changes?

The evaluations are intended to be a tool for quality assurance for the teachers, the department education boards and the university management. The results are summarized in histograms (for numerical scores) and free form text is aggregated. This information is used later on at three levels:

- The evaluations are available to the teacher before the last lecture where he/she is expected to present the results of the evaluation to the students. The teacher is also expected to present an action plan for the next time the course is held.

- All evaluations from a department are given to the department's study board, who looks at them and decides whether and which actions might be needed for some of the courses. The results are utilized differently across DTU departments. At some departments the standard overview is supplemented by using an average score of selected questions as an indicator of quality of the course or quality of teaching. At others a complicated somewhat ad-hoc aggregation of all questions is performed for each course and for each teacher.

- Finally, the deans of education also receive the evaluations for all courses at all departments and have the possibility of contacting the study boards, if something needs to be adjusted.

However, it is obvious that the evaluations contain much more information than is extracted now. For instance, data is actually collected and stored at the level of the individual student. By naive aggregation all information on both correlations between questions within each questionnaire and between answers of each individual student is totally ignored. Furthermore, other data might be included, e.g. the obtained exam grade, the student grade point average, student personal data (e.g. gender, nationality, age, previous grades, etc.) or course specific characteristics (e.g. course size, course workload, etc). Moreover, the information from student comments is now available only to the teachers of the course.

The aim of this thesis is to develop and apply statistical methods for the analysis of quantitative data from course evaluations (Form A and Form B) and to apply text-mining tools in order to extract information from open-ended quantitative student answers (Form C) and examine how it is related to the quantitative part of the evaluations (Form A and Form B).

Several studies on SET data investigate the relationship between student ratings and student achievements (Cohen, 1981; Feldman, 1989a; Abrami et al., 1997) and between student ratings and various student-specific, course-specific and instructor-specific characteristics Marsh (1987). However, it is also interesting to investigate the correlations between different SET questions. In particular, to find the degree of association between how students evaluate the course and how students evaluate the teacher. In this way it is possible to obtain a different angle on the perspective presented in work by Marsh (2007): that SETs primarily is a

function of the instructor rather than the course. As a subject we have chosen to study a single course over time.

A lot of research has been done on the use of multiple choice responses for course evaluations, but little investigation has focused on students' written comments, despite the fact that they are included in most of the surveys that colleges and universities use. However, students written feedback is as important as the students numerical answers. The students' comments provides better understanding of the meaning of the numbers from the quantitative part and can provide an answer for the question "why?" and can catch students' observations, recommendations, frustrations, satisfaction and any other issues that may not have been addressed in the numerical part of the survey.

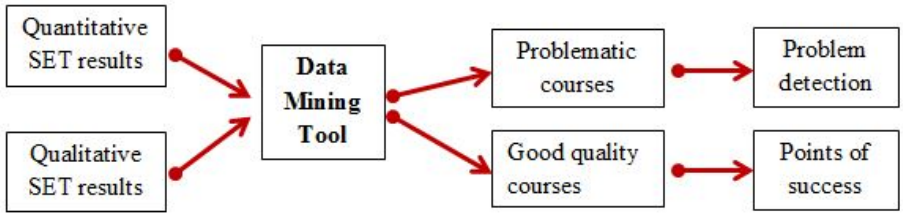
For teachers, the best way to analyse the students feedback is to read responses and picking up points of students satisfaction and/or dissatisfaction. Many teachers and course responsibles indicate that this section of the survey provides a more clear picture of what the students really feel or think. However, some authors are more sceptical and have pointed out that students are not trained observers and have little knowledge of formal evaluation of teaching (Braskamp and Pieper, 1981). Another problem of student comments is that the obtained response rates are usually much lower, than the response rates for the numerical part of the survey.

For university administration it is also important to know why SET results are low or high during analysis of SET results. However, it is too time consuming for administration to read trough all the student comments for all the courses of the university. In this situation, an automated method of extraction of the most important information from students written feedback may be able to provide insight to university administration and departments study boards on how a course was conducted, what went well, and what could be improved.

Figure 1.1 shows a possible scenario of automated analysis of student evaluations of courses and teaching quality. The results of both quantitative and qualitative feedback are passed to some data mining tool, where both open ended feedback and quantitative scores are analyzed. Based on this analysis courses could be divided into two groups of "Good quality courses" and "Problematic courses". Further analysis of problematic courses can be done to detect the points of student dissatisfaction, for example textbook or teaching methods. In this way the department study boards can focus on discussions of problematic courses. In addition, the evaluations of good quality courses can be analyzed in order to find what makes these courses good.

Varying policies exists on how and when student evaluations of courses and teachers should be performed: in the middle of the semester or at the end,

Figure 1.1: Illustration of automated analysis of the results of student course evaluations



before or after the final exam, before or after the students get the course grades. If the evaluation is conducted in the middle of the semester, the teacher can use the results in order to make some adjustments for the second part of the semester. The current group of students will benefit from such adjustments, not just for future students of the course. However, the mid-term SET can not provide the evaluation of the whole course. Is of interest to test statistically whether midterm evaluations can lead to improvement within the semester to meet the needs of the current course students.

While collecting the data from the mid-term evaluations at DTU, it was found that different groups of students participated at SETs at the middle of the semester and at the end of the semester. Students participation in SET is also highly debatable. If students who participate in SETs are dissatisfied with the course and the instruction, the results will be biased downward, but if students who are satisfied with the course are more likely to participate in SET, the results will be biased upward. Therefore, for both teachers and administrators it is very important to understand what kind of student's opinion is presented by the course evaluation results. If the bias exists, it should be taken into account, when analysing SET results. Moreover, such investigation of which course and students characteristics effect students SET participation, may provide an information about which group of students should be additionally encouraged to participate in SET.

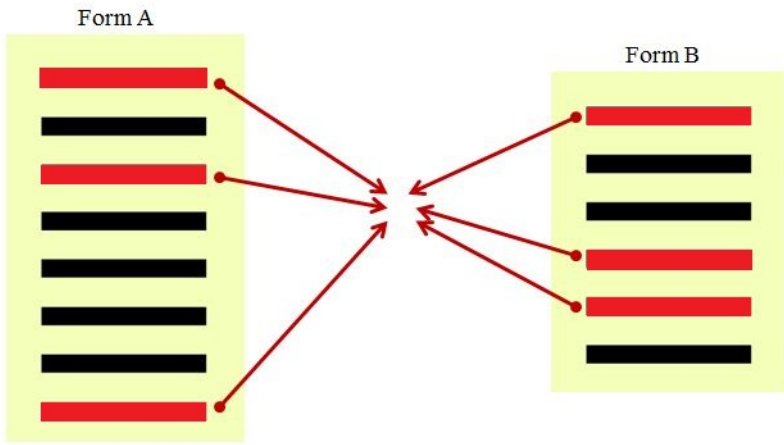
The goal of the current work is to develop efficient tools that will help to a better understanding of the results of the course evaluations at the Technical University of Denmark.

1.4 Thesis Objectives

The overall goal of the thesis is to investigate and apply statistical methods to the results of course evaluations in order to partially or fully answer the following questions:

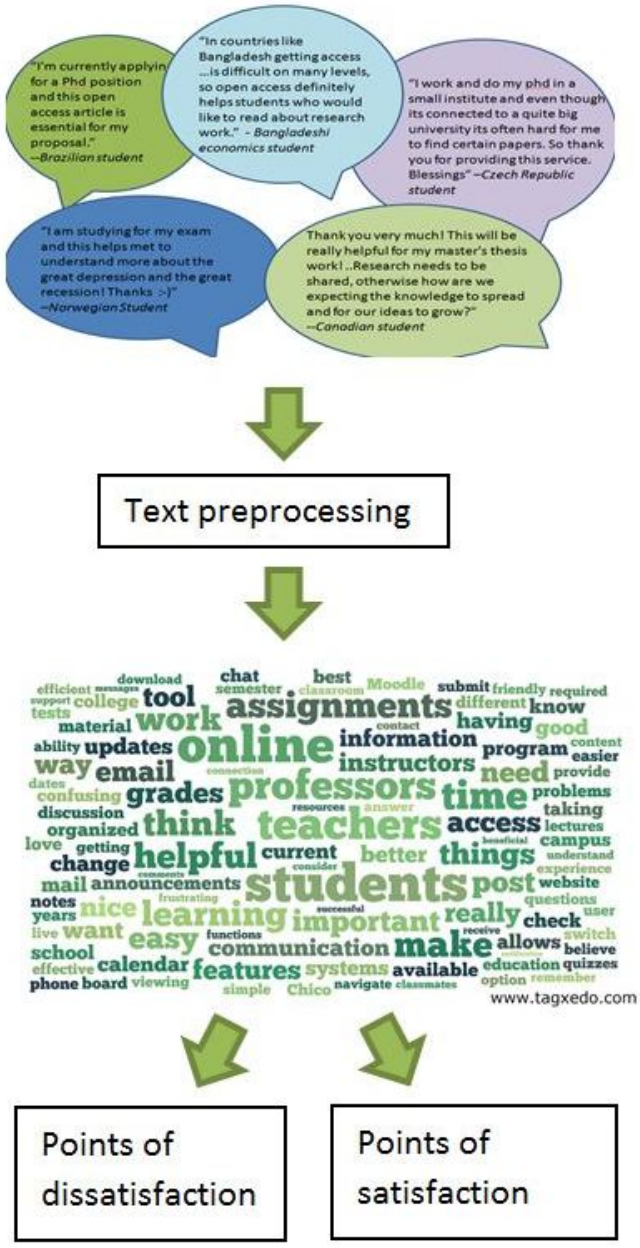
1. Is there a correlation and, if so, what is the structure of that correlation between Form A (evaluation of course quality) and Form B (evaluation of the teacher)? Figure 1.2 provides the illustration of a possible correlation structure. Some questions have more substantial contributions to the total correlations than others. If the structure of the correlation is similar for different courses or for the same course during some period, it might be beneficial to reduce the number of questions in the questionnaire. This can possibly lead to higher overall response rates or higher response rates of each question, since students tend to skip some questions.

Figure 1.2: Canonical correlation analysis illustration



2. Which methods can be applied to find the most interpretable model of association between two quantitative parts of evaluations: the evaluation of the course (Form A) and the evaluation of the teacher (Form B)? Data from student evaluations is characterized by high correlations between the variables. Moreover, the response rates on course evaluation at DTU are around 50% or lower, therefore the reliability of statistical tools used for SETs of small courses is questionable.
3. It is obvious that students' answers to open-ended questions provide more precise inputs to their satisfaction or dissatisfaction with the course or the

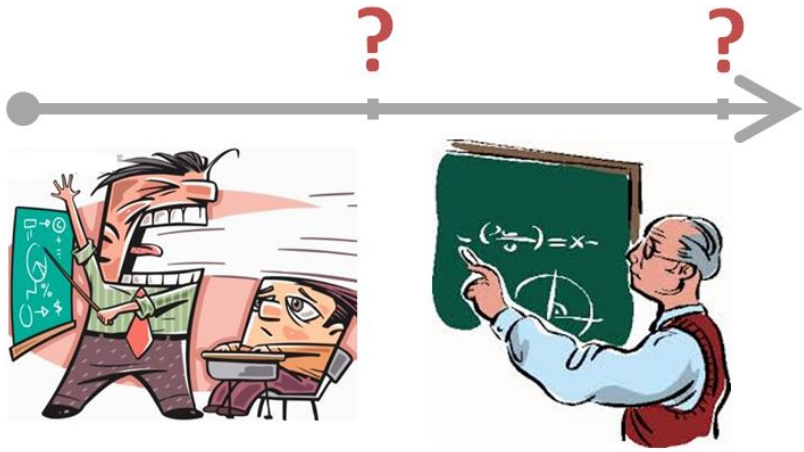
Figure 1.3: Text mining in student comments



teacher, than just a quantitative score. Figure 1.3 illustrates how text-mining methods can be used to extract aspects of student satisfaction or dissatisfaction. It is also interesting to examine the relationships between the information extracted from open-ended student comments and quantitative scores of SETs.

4. There are different opinions on when the student ratings should take place: in the middle of the semester or at the end, before or after the final exam, before or after the students get the course grades. There are advantages and disadvantages for all the mentioned settings. For example, if the SET took place in the middle of the semester, the teacher can react and make some adjustments suitable to the current group of students (Figure 1.4), not just for future students. On the other hand, the mid-term evaluation only provides information about the first part of the course, not the whole course. We wish to answer the question: what is the effect of mid-term course evaluations on student satisfaction with the course as measured by end-of-term evaluations?

Figure 1.4: Mid-term illustration



5. Another debatable issue of SETs is student non-response on the evaluation. Which student specific characteristics can be the determinants of whether or not a student participates in the evaluation? Are the students who submit an evaluation in the middle of the semester different from those who submit the evaluations at the end of the term? In addition, the impact of students and course specific characteristics on the SET scores was investigated.

In order to answer these questions various samples of data have been extracted from DTUs intranet - CampusNet.

1.5 Thesis Overview

This thesis is composed of two parts:

Part I is an introductory part that includes an overview of the most important research on course evaluation (Chapter 2), a description of various data sets used during the project (Chapter 3), a description of the methods used (Chapter 4), a discussion of the results (Chapter 5) and conclusions (Chapter 6).

Part II includes a selection of papers written during the project period. The first three papers are based on the analysis of a single course (Chapters 7, 8, 9, 10), while the others are based on data collected from multiple courses taught at DTU (Chapters 11, 12). The data for these papers were collected in the fall 2010, during the mid-term experiment, described in section 3.3.

A brief description of the papers is presented below:

Chapter 7 presents the analysis of association between information obtained from the course evaluation form and information obtained from the teacher evaluation form. By employing canonical correlation analysis it was found that course and teacher evaluations are highly correlated, however, the structure of the canonical correlation is subject to change with changes in teaching methods from one year to another.

Chapter 8 provides deeper analysis of the association between how students evaluate the course and how students evaluate the teacher using canonical correlation analysis (CCA). Data from student evaluations is characterized by high correlations between the variables (survey questions) within each set of the variables, therefore two newly developed modifications of the CCA methodology: regularized CCA and sparse CCA, together with classical CCA were applied to find the most interpretable model of association between two evaluations. The association between how students evaluate the course and how students evaluate the teacher was found to be quite strong in all three cases. However, applications of regularized

and sparse CCA to the present student evaluation data give results with increased interpretability over traditional CCA.

Chapter 9 analyses the positive written feedback on a well-established course in two subsequent years. The study tries to apply two clustering methods: k-means and SVD, to cluster the open-ended feedback. The study illustrates that many precise points of students satisfaction are reflected in student comments. Even for a well established introductory course, changes in teaching method can change the basis vectors for clustering comments. Moreover, study also illustrates problems, that arise when simple text-mining tools are applied to such short texts as students open-ended feedback.

Chapter 10 analyses the written responses to open-ended questions and their relationships with quantitative scores. A key-phrase extraction tool was used to find the main topics of students' comments, based on which the qualitative feedback was transformed into quantitative data for further statistical analysis. Application of factor analysis helped to remove outlier comments and to reveal the important issues and the structure of the data hidden in the students' written comments, while regression analysis showed that some of the revealed factors have a significant impact on how students rate a course.

Chapter 11 addresses how general university policies can influence the quality of courses by deciding when to perform student evaluations. To conduct the analysis an extra mid-term evaluation, identical to the final evaluation, was set up in the middle of the fall semester 2010 for 35 selected courses at DTU. The evaluations generally showed positive improvements over the semester for courses where the teacher had a access to results of the mid-term evaluation, and negative improvements for those without access. In particular, questions related to the student feeling that he/she learned a lot, a general satisfaction with the course, a good continuity of the teaching activities, and the teacher being good at communicating the subject show statistically significant differences.

Chapter 12 presents the investigation on the non-response bias at DTU and together with the investigations of whether student-specific and course-specific characteristics have an impact on SET scores. There was evidence of SET non-response bias both in the mid-term and end-of-term evaluations. Female students, with high GPA, taking the course for the first time were more likely to participate in the course evaluation survey at both time points. Analysis of SET scores showed that even though students with high GPA, had a higher probability to participate in the evaluation survey, the GPA itself had little effect on the SET scores. However, the grade obtained on the course was strongly positively related with both SET participation and SET scores.

Chapter 2

Analysis of student evaluation of teaching

The common means of evaluating teaching quality typically include course evaluation surveys, letters from students and colleagues, in-class evaluations, peer evaluations, alumni evaluations and teaching awards. However, one of the most widely used tools to assess classroom teaching is student evaluations of teaching or student ratings (Wright, 2006; Seldin, 1999; Centra, 1979)

Student evaluation of teaching and course quality is practised in many universities, schools and colleges around the world. SET is a very well documented and studied tool, however, its origin is relatively recent. It was first introduced in the literature by Kilpatrick in 1918, and more commonly adopted and studied in the 1930s and 1940s.

The period of the 1970s was a period of expansion of the research in course evaluation (Gravestock and Gregor-Greenleaf, 2008). Many issues of course evaluation have been discussed in the literature like validity and reliability of student ratings, faculty, administration and student perceptions of course evaluation, various relationships between student ratings and students, course and instructor characteristics, and other sources of potential biases in SETs. There are also different policies and practices regarding student ratings like the design of the SETs, the access to the results, the timing of the SETs, the interpretation

and use of the results, etc.

There are a number of meta-studies, that provide an overview of the SET issues: meta-studies and own research by Marsh and his co-authors (1981; 1982a; 1982b; 1984; 1987; 1991; 1992; 1992; 1997; 2000; 2007), Feldman (1978; 1979; 1989a; 1989b; 1993; 2007), McKeanie (1969; 1979; 1997), Cohen (1980; 1981), Wachtel (1998), Centra (1976; 1993; 2003; 2009), works by d'Apollonia and Abrami (1997; 1997; 1991; 2007), summaries of research by Cashin (1988; 1994; 1995), review of student evaluation practices by Gravestock and Gregor-Greenleaf (2008) and other works.

One of the most often-cited overviews of research on student ratings of instruction is the monograph by Marsh (1987). In 1982b Marsh, an internationally recognized expert in the area of psychometrics, developed Students' Evaluation of Education Quality (SEEQ) instrument. SEEQ has been extensively tested and used in more than 50,000 courses with over one million students at both the graduate and undergraduate levels. The meta-analysis (Marsh, 1987) provides an overview of findings and of research methodology used to study students' evaluations of teaching effectiveness and results of the author's own research. A study was updated in 2007. The papers demonstrates that SETs are:

- multidimensional;
- reliable and stable;
- primarily a function of the instructor who teaches a course rather than the course that is taught;
- relatively valid against a variety of indicators of effective teaching;
- relatively unaffected by a variety of variables hypothesized as potential biases;
- seen to be useful by faculty as feedback about their teaching, by students for use in course selection, and by administrators for use in personnel decisions.

This chapter gives an overview of the most discussed research questions of student course evaluations.

2.1 Reliability of course evaluations

The issue of reliability is of great concern in using student evaluations of instruction for making comparative decisions about faculty and courses. In education,

reliability usually refers to consistency, stability and generalizability of the measurements. Reliability of SETs is also dependent on the number of ratings, the more students participate in evaluations, the more reliable are the results.

Regarding student evaluations of course, reliability often concerns consistency, which means that within the same class students tend to give similar ratings to a given item. Most of the literature agrees that the SETs are reliable tools because they provide consistent and stable measures for specific items, like the instructor's skills or the course workload (Abrami, 2001; Hobson and Talbot, 2001; Wachtel, 1998). This is particularly true when the tool has been carefully constructed and psychometrically tested before use (Centra, 1993; Aleamoni, 1987; Marsh, 1984).

The question of stability mostly deals with agreement between rates over time. In general, ratings of the same instructor tends to be similar over time (Overall and Marsh, 1980; Centra, 1993; Braskamp and Ory, 1994). Overall and Marsh (1980) in his longitudinal study investigated the long-term stability of SETs by analysing 1374 undergraduate and graduate business administration majors from 100 classes who completed their programs at a comprehensive state university between 1974 and 1977. Results show large and significant correlations between current and retrospective SET results.

Generalizability refers to how accurately the data reflects what students think and how the teacher performs teaching, not just how effective teaching was at a particular course in a given term. Marsh (1982a) in his analysis of 329 classes have found that teachers ratings did not change significantly when teaching different courses. Gillmore and Kane and their co-authors 1978; 1976 showed that generalizability theory can be a valuable tool in analysing student evaluations. Gillmore et al. (1978) compared courses from instructors who had taught two different courses and courses that had been taught by two different instructors. When teachers were the objects of measurement, authors found generalizability coefficients to be quite satisfactory. Smith (1979) in a similar research settings found that generalizability coefficients were much higher for making decisions about instructors with instructor-related rather than course-related items. Marsh (1982a) in his analysis of 1324 courses concluded that the instructor, not the course, is the main determinant of students' ratings.

Generalizability is a very important issue when making personnel decisions. Keeping in mind such decisions should also be based on some additional information beyond the SETs.

2.2 Validity of course evaluations

Although, most of the researchers may agree that student evaluations of teaching are reliable tools, there is somewhat less consensus regarding their overall validity. Validity determines if what is supposed to be measured by SETs is being measured in reality. Researchers have tried different approaches, collecting data that either support or contest the conclusion that student ratings reflect the effectiveness of teaching.

A historical overview of the research by Greenwald (1997) notes that the majority of publications produced during the period 1975-1995 indicate that course evaluations are valid. McKeachie (1997) also concluded that the student course evaluations are the "single most valid source on teaching effectiveness".

Researchers have tried to compare the results of the student evaluations to different measures of student learning such as students actual or expected grades (Feldman, 1989a; Cohen, 1981) or to other evaluations of teacher effectiveness such as instructor self-evaluation (Feldman, 1989a; Marsh and Dunkin, 1992), ratings by administrators (Kulik and McKeachie, 1975; Feldman, 1989b), peer evaluations (Kulik and McKeachie, 1975; Feldman, 1989b), and alumni evaluations (Overall and Marsh, 1980; Braskamp and Ory, 1994; Hobson and Talbot, 2001).

There are various ways to asses student learning. The obtained grade is one of the most popular measure of students learning used in literature. There are a number of studies which tried to compare multiple-section courses, the courses where different parts of the course are performed by different teachers using the same syllabus and textbook and external exam. Cohen (1981) and Feldman (1989a) made a review of these studies. The reported relationships confirmed the validity of SET, the classes where teachers get higher SET scorer students also scored higher on the external exam (Cashin, 1995). More discussion on the relationship between the student ratings and grades is presented in section 2.6.

Instructor self-evaluation is another tool used to establish validity of SETs. Feldman (1989b) provides a list of 19 papers that investigated the correlation between student evaluations and instructor self-evaluations. The average correlation is 0.29.

Feldman (1989b) provided a review of the research comparing the SETs made by current and former students, colleagues, administrators, external (neutral) observers, and the teachers themselves. The author reports the average correlation between student ratings and ratings by administration to be 0.39 (based on results of 11 studies) and between students ratings and peer-ratings to be

0.55 (based on results of 14 papers).

To address the issue that the current students may not adequately judge the long-term effects of instruction some studies checked how the student ratings are related to the retrospective ratings of the same instructor provided by the same students several years later. Different studies report such a correlation to be quite high. Feldman (1989b) reports an average correlation to be 0.69, based on results of 6 cross-sectional studies. The study by Overall and Marsh (1980) showed large and statistically significant correlations between end-of-term and retrospective ratings. The authors analyzed the data of more than 1,000 undergraduate and graduate business administration students from different classes that provided feedback at the end of each class, and again at least one year after program completion.

Some studies used external observers, who were trained to make classroom observations. Feldman (1989b) in his meta-study reported an average correlation of 0.50 between the student ratings and ratings of trained observers. Kulik (2001) in his review mentions a careful study by Murray (1983), where 6 to 8 trained observers visited classes taught by 54 university lecturers receiving either low, medium, or high student ratings in an earlier semester. Highly rated teachers tend to receive high scores and low-rated teachers tend to receive low scores from observers, especially in such teaching qualities as clarity, enthusiasm, and rapport.

2.3 Online-based vs paper-based course evaluations

The traditional way to obtain student evaluations of the course and the teacher is to distribute printed questionnaires and survey forms among students at the end of the course, while more modern techniques are based on on-line questionnaires.

The in-class paper-based method of conducting evaluations is less likely to suffer from the effects of non-response, because most of the active students are assumed to be in class when evaluations are conducted. However many universities have switched to the online-based student evaluations, that offer several advantages over paper-and-pencil evaluations.

For the web-based course questionnaires students can respond outside of class at their convenience (Dommeyer et al., 2002; Layne et al., 1999), therefore it requires less class time. Additionally, web-based questionnaires provides a less

expensive method of collecting course evaluation results and can provide results immediately

A literature review by Anderson et al. (2005) suggest that there is some evidence that the Web-based evaluation methods lead to lower response rates. Hoverer, study by (Avery et al., 2006), that analysed SET results conducted on-line or on paper of 29 courses taught between 1998 and 2001, and study by Fike et al. (2010), that employs the sample of student evaluations of 247 courses, showed that there is no evidence that evaluation scores change when evaluations are completed online rather than on paper. Lower response rates may occur for several reasons: students concern about anonymity, computer technical difficulties, and the time required to respond outside of class.

Dommeyer et al. (2002) analysed settings where sixteen professors who taught two sections of the same class and were randomly assigned to have one of their sections evaluated by the in-class method and the other by the on-line method. The traditional paper based method had a higher response rate than the web-based method. During the post-evaluation survey the on-line responders complained that web-based evaluation process may not be anonymous and that the log-on system was time-consuming. Analysis of almost 2500 students, who were randomly assigned to either the traditional or the electronic evaluation by Layne et al. (1999) also showed that students were more likely to evaluate their teachers when the evaluations were conducted in class.

Additionally, Layne et al. (1999) found that average ratings did not differ between the two methods of conducting SET. This fact is also confirmed by further investigation by Dommeyer et al. (2004) and Donovan et al. (2006), who analysed 11 courses with settings similar to Dommeyer et al. (2002).

Donovan et al. (2006) found differences between the two methods of conducting SETs in number and length of comments. Students completing faculty evaluations online wrote more comments, and the comments were more often formative in nature. Layne et al. (1999) also found that the response rates to open-ended questions posted on-line tend to be higher. Hardy (2003) examined 26 classes in which the same instructor taught the same class multiple times. The classes using paper rating forms which were compared with the classes evaluated online. The overall response rate in the classes evaluated online was lower. However, students wrote many more comments and students who respond online write more detailed comments. Similar findings are also presented in works by Anderson et al. (2005); Johnson (2003a); Kasiar et al. (2002); Ballantyne (2000).

Crews and Curtis (2011) analysed the data gathered from university faculty, who transitioned from traditional paper to online course evaluations. Authors provides some suggestions for universities transitioning from traditional paper-

based course evaluations to web-based course evaluations. One of the most important suggestions is providing adequate training for faculty members on the on-line course evaluation system, ensuring that when students withdraw from a course that they are dropped from the evaluation system, sharing with faculty strategies to increase the student response rate.

Norris and Conn (2005) suggested that usage of reminder e-mails from instructors or university administration and messages posted on on-line class discussion boards and forums can increase response rates. Dommeyer et al. (2004) showed that when a grade incentive (one quarter of one percent for any student who had completed the on-line evaluation) was used to encourage response to the online survey, a response rate was achieved that was comparable with that to the in-class survey.

Introduction of the web-based evaluation systems had a positive effect on the way the results are summarized and used. It enabled an easy comparison of the courses within each department or university/college. Overall, theoretical and practical considerations, seem to lead to the conclusion that the online-based SETs bring remarkable advantages. The cost reductions, time savings and the ease of calculation of results, overcome the disadvantages of web-based evaluation surveys.

2.4 Faculty, Administrator and Student Perceptions of Course Evaluations

Some studies have been conducted on the attitudes and perceptions about course evaluation systems by those who use them and who are affected by them: faculty members, administrations and students.

Many **faculty** members are suspicious of student evaluation of course and teaching quality mainly due to the belief that students are not competent enough to give appropriate evaluation (Nasser and Fresko, 2002; Ryan et al., 1980). However, some more recent studies (Harun et al., 2013; Smith and Welicker-Pollak, 2008), showed that the lecturers, agreed that students have the right to judge the quality of the teaching, but doubt the accuracy of the ratings.

Another teachers concern is that the student's grade expectation may influence the student ratings (Baldwin and Blattner, 2003; Nowell, 2007). The instructors negative perceptions of evaluations can lead to ignorance of the importance of SETs and can become an obstacle in the way of teaching and course improvement efforts.

Beran and colleagues in their studies (2005; 2007; 2009) investigated the utility of SETs for students, faculty, and administrators. The results shows that the majority of faculty believes that the evaluation data was being used appropriately by academic administrators. The study revealed that a majority of the instructors surveyed generally had positive views of course evaluation. (Beran and Rokosh, 2009) consisted of 357 instructors attitudes towards student ratings. The teachers tend to agree that the student rating useful to administrators in making summative decisions.

The most common **administrative** use of evaluation data is personnel decisions. (Gravestock and Gregor-Greenleaf, 2008) in their review of research on student evaluation state that most studies showed that university and college administrations have a positive attitude toward student ratings and find them a useful source of information.

Research on **students** perceptions of SETs is limited Gravestock and Gregor-Greenleaf (2008). Some small studies showed that students think that the process of collecting students feedback is useful and that students are valid evaluators of teaching.

There is evidence that students feel that evaluations have no effect on teacher performance and believe that faculty and administrators don't take their evaluations seriously (meta-study by Wachtel (1998)). Moreover, students do not know if anyone other than the instructor sees the evaluations nor understand that ratings have an impact on personnel decisions. However, a study by Gaillard et al. (2006) demonstrated, based on a sample of 389 students, that students are more likely to complete course evaluations if they understand how the evaluations are being used and believe that their opinions have an effect.

Paper by Beran et al. (2005) demonstrates that many students make little use of SET data: it was discovered that 56% of students did not use rating data at all. Of those students who indicated they used the evaluation results, less than $1/3$ used them to select courses based on content and structure and almost $2/3$ used them to select courses based on the instructor.

2.5 Students, Course and Instructor Characteristics and their Impact on Course Evaluations

In response to ongoing concerns about the validity of the student evaluations, many researchers have investigated whether factors unrelated to teaching skills

and course structure can explain the variability in ratings. There is much evidence that extraneous factors such as student characteristics, course characteristics or other environmental characteristics may influence how a student rates courses or/and teachers.

Students are very diverse and some **student specific attributes** can significantly affect how they rate their instructors and courses. These characteristics can include gender, race, age, nationality, cultural/ethnic background, academic major, motivation for taking a course, obtained grade, grade expectations, student grade point averages (GPA), learning style, knowledge of prerequisite material, years in school and students' interest in the subject matter prior to enrolling in the class .

Table 2.1 provides an overview of relationships found between student ratings and student-specific characteristics.

Table 2.1: Overview of relationships found between student ratings and student-specific characteristics.

Characteristic	Summary of findings
Motivation or prior interest	Student with higher interest in the course tend to evaluate these course more favorably (Cashin, 1988, 1995). However, it is not always clear if interest existed before start of course or was generated by teacher. Majors tend to rate instructors more positively than non-majors (Feldman, 1978).
Expected / obtained grade	Aleamoni (1999) in his meta-study has identified 37 studies that revealed positive significant correlations between expected/obtained grades and SETs, and 24 studies that found this kind of relationship being insignificant
Gender of student	Gender of students is not related to his or her responses, however Students tend to give slightly higher ratings to teachers of the same gender (review by Ory (2001)).
Level of student	Weather student is master or bachelor has little or no effect on ratings (McKeachie, 1979).
Student GPA	Davis (2009) in a summary of research, cited several studies that shows little or no relationship between student ratings and GPA.
Age	Most of the studies report little or no effect of students agr on student ratings (McKeachie, 1979; Centra, 1993).

The mood of a student at the time he or she answers the questions of the evaluation questionnaire is another factor outside the control of the instructor, and this may also have an impact on how the students rate a course and teacher. Munz and Munz (1997) based on sample of 136 students found that student

positive mood state at the time of the evaluation accounted for only a modest 4% to 6% of the variance in the instructor and course ratings. However, LaForge (2003) did not find the correlation between instructor effectiveness ratings and students mood measures to be significant, based on survey of 241 students.

The **instructors characteristics** like race, gender, rank, experience, weight and dress may also influence the rate students give to their teachers.

In general, instructor age, rank and experience are not correlated with student ratings (Cashin, 1995). Marsh and Hocevar (1991) in their work made a longitudinal study of student evaluations of the same teachers across 13 years and found no systematic changes within teachers over time.

Instructor's entertainment level may positively influence the student ratings of the instruction. Study by Naftulin et al. (1973), in his "Dr. Fox" study concluded that teachers enthusiasm can influence student rating. These results raised doubts about the usefulness of teaching evaluation. Felton et al. (2004, 2008) found in their studies of the web page `http://RatemyProfessors.com` that students gave higher ratings to the instructors they deemed "hot" or good looking, and not to the most helpful teacher or instructor, whom they learned the most from. However, a study by Marsh and Ware (1992) showed that when students have incentives to learn, the entertainment level of an instructor is less important.

Different **course characteristics** like the size of the class, course difficulty, course workload, whether the course is mandatory or elective, level of the classand time of day of the class may influence student evaluations. In addition, some environmental characteristics like physical attributes and the ambiance of the classroom should also be considered. Table 2.2 provides a summary of findings on relationships between various course characteristics and course evaluations.

To be fair to teachers, the effects of extraneous factors on student ratings, that are out of teachers control, should be investigated and properly dealt with.

2.6 Grade Inflation and Student Ratings of Instructors

In the literature regarding course evaluations of teaching, special attention is given to the relationships between students' grades, both actual and expected, and student ratings of the instructor. So called grade inflation occurs when

Table 2.2: Overview of relationships found between student ratings and course characteristics.

Characteristic	Summary of findings
Workload / difficulty	Many teachers have a belief that harder courses results in lower evaluations. However, some studies found that more difficult courses requiring more effort and time tend to receive somewhat more favourably ratings (Marsh, 2007; Cashin, 1988; Centra, 1993)
Class size	Some of the studies have found that smaller classes get lightly higher evaluation ratings (Centra and Creech, 1976). While other researchers did not found statistically significant correlation between the course size and the course ratings (Marsh and Roche, 1997). Centra (2009) found that smaller classes not only tend to receive higher ratings but that students in those classes report learning more. It is not clear whether this reflects differences in teaching methods typically used in the two contexts, or whether it is an effect of size alone. McKeachie (1997) and number of other researchers suggests that it is not accurate to compare SET results of courses with large difference in number of students.
Level of course	Graduate level courses rated somewhat more favourably (Cashin, 1995; Marsh and Roche, 1997).
Mandatory / elective	Marsh (2007) in his meta-study concludes that courses and those with higher percentage taking course for general interest tend to get higher ratings than required courses. While some research works found that whether course is mandatory or elective have no statistically significant impact on SET ratings (Cashin, 1988, 1995).
Academic discipline	Some studies have shown that particular disciplines receive higher ratings. Humanities and social sciences courses tend to get higher SET scores then the natural sciences courses (Gravestock and Gregor-Greenleaf, 2008; Wachtel, 1998). This reflects the difference in teaching styles used. The course evaluation comparisons of courses from different disciplines is not appropriate (Cashin, 1988, 1995).

higher grades are given for the work that would have received lower grades in the past.

Many researchers are concerned that students give higher evaluation scores to those teachers who reward them with good grades. Student course evaluations are often used by educational institution administration in personnel decisions,

like promotion and tenure. A teacher may therefore improve evaluations by improving their teaching, or by awarding students with higher grades for assignments and exams.

Marsh and Dunkin (1992) hypothesized that the grades/ratings correlation could be a result of instructors setting lower grading standards with the purpose of receiving better evaluations. Some research results support the hypothesis that instructors who are "easy", with lower grading standards, and not necessarily the best teachers tend to receive higher ratings An alternative hypothesis is that more effective instructors motivate students to work harder, learn more, and therefore earn better grades.

A comprehensive study by Johnson (2003b) showed a statistical correlation between high grades and high course evaluations. Weinberg et al. (2007) conducted an analysis of about 50,000 student evaluations in 400 economics courses over a period of several years. The paper showed that the student evaluations were positively related to the current grades but unrelated to learning, which would affect future grades.

The relationship between the student expected grades and the student evaluations of teaching is also controversial. Chacko (1983) analysed two groups of students: in one of the groups the mid-tem exams were more harshly graded than in the other group. The results of course evaluation show that the ratings in the experimental group were significantly lower than in control group. However, Seiver (1983) found that when a two-stage least squares (2SLS) procedure was employed to control for the endogeneity, there is no evidence that instructors are inflating grades in order to better their SET scores.

One of the criticisms among opponents of SET's is that student expectations are not under control of the instructor, and therefore, the bias skews teacher assessment results. Centra (2003), in a study that involved over 50000 students, concluded that expected grade generally had no effect on ratings of teachers and courses. Cashin (1995), in his review of research, lists a number of papers that found positive but low (0.1-0.3) correlations between student ratings and student grade expectations. Isely and Singh (2005) found the opposite conclusion: if an instructor of a particular course has some classes in which students expect higher grades, a more favourable average rating of the instructor is obtained in these classes. The authors used class-specific data, rather than course-specific, from 260 classes and controlled for course differences by employing a fixed-effects model.

2.7 Timing of Course Evaluations

There are various policies regarding the time, when the course evaluation should take place: in the middle of the semester or at the end, before or after the final exam. The most common policy among universities is to conduct the course evaluation at the end of a particular unit of instruction (Gravestock and Gregor-Greenleaf, 2008). Abrami et al. (2007) pointed out that the timing and the method of collecting student ratings also has effect on the results of the evaluation.

Some researches investigated the timing of the evaluations - midterm, end-of-term, before or after the exam, but the number of studies that investigates specifically the timing of evaluations is limited.

Frey (1976) did not find a statistically significant difference between students ratings of the courses, where half of the students submitted their evaluation during the last week of the course (before the exam) and the other group submitted their evaluations during the first week of the subsequent semester (after final exam). Moreover, the analysis of the relationship between final exam grade and student evaluation was found to be strong in both cases.

Cashin (1995) summarized that the timing of the evaluations was not significant. While Feldman (1979) in his review of existed research concluded that evaluations administered at any time during the second half of the term seemed to yield similar ratings.

Students often express the opinion, that even if the instructor is interested in their feedback, it will not have an effect on this semester, and thus current students will not benefit from giving feedback. Midterm student feedback has great value in higher education for a variety of reasons. First, it provides an overview of student perceptions about a course while course changes that semester are still possible. Second, the same students that provide the midterm feedback can benefit from course improvements. On the other hand end-of-semester evaluations summarize students' overall satisfaction or dissatisfaction with the course, but at the point when it is too late to make adjustments in teaching in the current semester. Asking students to provide feedback at the middle of the course makes it more clear that the teacher takes evaluation results seriously.

A study by Overall and Marsh (1979) analysed 931 student who were asked to complete 3 surveys: an pretest survey (about grade and course expectations), a midterm survey and an end-of-term survey. Authors concluded that the students feedback collected at midterm results in: more favourable student ratings at the end of the term, better final examination scores, and more favourable affective

outcomes. Cohen (1980) in his meta-analysis concluded that on average the mid-term evaluations had made a modest but significant contribution to the improvement of teaching. These findings suggest that feedback from student ratings, particularly when coupled with a discussion of their implications, can be an effective tool for improving teaching effectiveness.

2.8 Non-response in Course Evaluations

As described in section 2.5, there is evidence that student evaluations of teaching may be influenced by student-specific characteristics, instructor-specific characteristics, course-specific characteristics and also environmental characteristics. However these characteristics can also have an impact on whether the students respond to the evaluation questionnaires or not. Moreover, a non-response bias may also occur when a participant does not complete part of the evaluation survey. Since the results of SETs are often used in personnel and administrative decisions, non-response deserves considerable attention.

A lot of research papers state that the SET response rates are dependent on the method of administration, paper based vs. online based. Due to migration of survey mode from traditional paper and pencil to web-based instruments , the migration generally leads to a decrease in response rates (see section 2.3). Despite this fact, more and more universities move towards on-line methods of evaluation administration, that are cheaper to conduct and that can provide results immediately. Moreover, on-line student responses can be coupled with student specific characteristics in order to investigate what kind of students are more likely to participate in student evaluations.

Some researchers (Cohen, 1981; Costin, 1978; Isely and Singh, 2005; Marsh, 2007) have found that high achievers tend to rate their instructors more favourably. On the other hand, there is also an evidence that the students who earned higher grades on a course and students with higher cumulative GPA are more likely to fill the evaluation forms, while students who are doing poorly in a course should be less expected to submit course evaluations or rating questionnaires (Avery et al., 2006).

There is some evidence that female students are more likely to evaluate than male found in studies by Avery et al. (2006), Kherfi (2011) and other studies. Kherfi (2011) in his analysis of more than 4,000 students from 376 courses, also found that freshmen or new college students are more likely to evaluate the course. This can be explained by the fact that first year students are more enthusiastic about university life, and have higher hopes about how their as-

sessments of their professors can make a difference. However, the trend shows that as students go on and get more used to college life, some of them think that evaluations are not taken seriously by the university.

There have been three recent dissertations on nonresponse to students ratings. Jones (2009) incorporated nine variables to determine nonresponse and found that gender, ethnicity, and final course grade were determinants of non participation in a survey. Fidelman (2007) examined both undergraduate and graduate students at Boston College and found that gender, expected grade, year in school, and teaching experience were predictors of nonresponse. Both studies were also focused on predictors of how students rate courses and instructors. Adams (2010) focused only on nonresponse for online based SETs, using a large dataset that incorporated more variables, more students, and more course evaluations. It was found that characteristics of the course also tended to influence students participation in the survey (Adams, 2010; Adams and Umbach, 2012). Submissions of SETs were more likely when the course and the students major were in the same department. They probably think that courses in their own major were more important than all other courses.

Porter and Umbach (2006) analysed survey data from 321 institutions and found that institutional characteristics such as public/private status and urban location affects response rates.

Overall, there is evidence that the results of evaluations can be biased. However, course ratings are still considered a reliable source of information on teaching effectiveness because many universities now use different strategies and approaches to extract real and honest feedback from their students.

2.9 Analysis of Students Open-ended Comments

There is a great range of student evaluation forms currently being used by different educational institutions around the world. Most of the student evaluation survey questionnaires fall into three general categories : a purely quantitative questionnaire, a purely open-ended questionnaire or a combination of the quantitative scales and qualitative questions.

Open-ended questions have much lower response rates than the quantitative questions. Completion of the written open-ended questions are lower when course evaluation is conducted in the traditional paper and pencil way, while the response rates on the open-ended questions are much higher, when the evaluation is done using computer-based questionnaires (Aleamoni, 1987; Abrami

et al., 2007).

Students written comments have not received as much attention as the quantitative data from student evaluations. Analysis of the open-ended students' comments is problematic, since they have no built-in structure and can range from just a few words to paragraphs of detailed analysis of positive and negative issues of a course, teacher and teaching material. In general, students more often write positive comments, rather then negative, and comments tend to be more general rather than specific (Alhija and Fresko, 2009). Some faculty members give preference to students written comments over open-ended questions over the rating scales (Cashin, 1995). This is mostly because the open ended nature of a question allows students to focus on what exactly is most important for them. Many faculty members consider written comments as more credible for the purpose of self-improvement.

Studies on analysis of written comments, that have been published, suggest how written student comments can be organized and analyzed in order to reveal information about aspects of the learning process. Most such studies suggest manual categorization of comments into groups of positive, negative and neutral, or some other kind of grouping, with further investigation of particular factors that reflect students satisfaction or dissatisfaction within each group. Lewis (2001) in his paper discusses how qualitative research techniques can be applied to the analysis of student written comments. The author suggests to start the analysis by classifying the student comments according to students overall course satisfaction and then adding another dimension, that can show where changes might be made. These additional dimension can be different components of effective teaching, fair grading or other issues.

Not much research has been done to investigate the relationship between the content of written comments and data obtained from the quantitative part of evaluations. Ory et al. (1980) and Braskamp and Pieper (1981) in their studies found that students generally provide similar evaluations of course and instructor quality on both open-ended and numeric questions. The first paper found the correlation between a global instructor item and students' written comments to be 0.93 (a sample of 14 classes was used) and in the second paper the correlation was found to be 0.75 (a sample of 60 classes was used). More recent work by Burdsal and Harrison (2008) also found a strong positive correlation between students comments and students ratings in their analysis of 208 classes. These studies suggested that the information from student ratings considerably overlaps the information in student comments.

Improvement of computational power and the development of text mining methods allows for a more sophisticated analysis on teacher and course evaluation data. However, studies that apply text-mining tools to analyse students feed-

back are relatively rare. A recently published dissertation on text mining in students comments by (Jordan, 2011) suggests that the student comments are moderately related to the quantitative scores from course evaluations. Moreover, some patterns found in student comments provided additional information that was not revealed by analysis of quantitative scores.

2.10 Scandinavian Studies on Course Evaluations

Most of the research in the field of course evaluation of teaching has been done based on North American datasets. The data collected by different universities, colleges and schools in USA is usually open to the public and have a high quality. Studies based on data from universities outside USA and Canada are scarce.

In Denmark as in other Nordic countries, the general use of course evaluations has a shorter history. SETs have primarily been introduced for formative purposes as well as an instrument for the institution to monitor and react on student satisfaction in general and on specific issues. As an effect of a requirement from 2003, all Danish universities make the outcome of course evaluations public (Andersen et al., 2009). Thus, key results of the existing SET processes are also used to provide information to students prior to course selections.

Johannessen et al. (1997) conducted a study based on a sample of Norwegian high-school students to retrieve evaluative dimensions of their teachers, i.e. to gain insights into what students emphasize in their evaluation of teachers.

Among studies of university student evaluation of teaching, Westerlund (2008) investigated the effect of class size on student ratings for the introductory mathematics course at Lund University in Sweden. The impact of class size was found to have a significant negative effect on the perceived quality of the course.

Pulkka and Niemivirta (2012) examined the relationships between student achievements, course evaluations and performance using the data collected from the Finnish National Defence University. The results suggested that performance and student course evaluations were to some extent influenced by goal orientations and by different pedagogical practices.

2.11 Other Issues of Student Evaluations

The present chapter discussed various issues of course evaluations, that can be statistically analysed. In addition, there are a number of non-statistical papers that investigates various important problems of SETs. The most important are:

- Access to the results of course evaluations. Who get the access to SET results and when? For example, Haskell (1997) suggests that teachers, who get the evaluation results before the last lecture, can benefit from in-class discussion of SET results with the students.
- Effects on course selection of allowing students to see the results of course evaluation. Students are more likely to select a course or teacher that has higher rates over the lower rated course/teacher (Wilhelm, 2004; Coleman and McKeachie, 1981).
- Importance of confidentiality and anonymity of student ratings. Most of the researchers recommend that SET should be anonymous (Cashin, 1995). However some suggested that the institution, not the instructor, should be able to identify who participated in student ratings (Wright, 2006). This will allow administrators to make some follow-up investigations. For example, check where extremely high or extremely low scores comes from.

2.12 Literature Summary

Student evaluations of teaching have been widely criticized, especially by teachers, for not being accurate measures of teaching effectiveness Survey data collected for the purpose of evaluation is perceived differently by faculty, students and university/school administrations. Many teachers argue that factors outside of their control like student grade expectations, have an effect on students ratings.

Despite this more than 90 years of validity and reliability research supports the value of student course surveys as a method of data collection on student satisfaction with teaching. The timing of evaluation also has an effect on the survey results. The evaluations conducted in the middle of a semester can demonstrate greater student participation than a survey conducted at the end of the semester as students may feel that they may have an impact on their own course.

Additionally, student written comments may provide useful and actionable information beyond what can be learned from the standard Likert scale questions. With the help of text mining tools the unstructured data from open-ended feedback can be accessed and/or integrated into analysis of SET results on the institutional level.

Chapter 3

Data

There is a great range of different student evaluation forms currently being used by the different educational institutions around the world. However, certain elements are almost universal. Course evaluations should be anonymous and are most commonly distributed at the end of a particular unit of instruction (Gravestock and Gregor-Greenleaf, 2008). Most of the student evaluation surveys fall into three general categories (Sheehan and DuPrey, 1999; Alhija and Fresko, 2009):

- a purely quantitative questionnaire, primarily of Likert scale questions (Likert, 1932).
- a purely open-ended questionnaire.
- a combination of the quantitative scales and open-ended questions, which is most frequently used

Course evaluation surveys generally include questions about communication skills, organizational skills, enthusiasm, flexibility, attitude toward the student, teacher - student interaction, encouragement of the student, knowledge of the subject, clarity of presentation, course difficulty, fairness of grading and exams, and global student rating. Student interviews can also be a useful method of data collection (Abbott et al., 1990).

3.1 The Course Evaluation System at DTU

Teacher evaluations and overall course quality evaluations are widely used in higher education around the world. Students submit their feedback about the teacher and the course anonymously during the course or at the end of the course, before or after getting the final grade. The results of evaluation are usually employed by the teacher and/or by university management to improve courses for future students and to improve instructor effectiveness.

The traditional way to obtain student evaluation of the course and the teacher is to distribute printed questionnaires and survey forms among students at the end of the course, while more modern techniques are based on on-line questionnaires. Many universities are switching from paper to web-based SETs to decrease costs and facilitate the mode of data collection and analysis.

At the Technical University of Denmark (DTU), as in many other universities around the world, students regularly evaluate courses. Since 2001 standard student evaluations at the DTU have been performed using web-based questionnaires posted on "CampusNet" (the university intra-net) in the last week of the semester, before the exams and the grades are given. The evaluation form consist of tree parts:

- Form A contains specific quantitative questions about the course (Table 3.1)
- Form B contains specific quantitative questions about each teacher of the course (Table 3.2)
- Form C gives the possibility of more qualitative answers on 3 questions:

  C.1.1 What went well?

  C.1.2 What did not go so well?

  C.1.3 Suggestions for changes?

The students rate the quantitative questions on a 5 point Likert scale (Likert, 1932) from 5 to 1, where 5 means that the student strongly agrees with the underlying statement and 1 means that the student strongly disagrees with the underlying statement. For question A.1.6 5 corresponds to "much less" and 1 to "much more", while for question A.1.7, 5 corresponds to "too low" and 1 to "too high". Question A.2.1 is active only for courses where English is the main language of teaching (all master-level courses at DTU plus some bachelor courses).

Table 3.1: Questions in Form A

	Question
A.1.1	I think I am learning a lot in this course
A.1.2	I think the teaching method encourages my active participation
A.1.3	I think the teaching material is good
A.1.4	I think that throughout the course, the teacher has clearly communicated to me where I stand academically
A.1.5	I think the teacher creates good continuity between the different teaching activities
A.1.6	5 points is equivalent to 9 hours per week. I think my performance during the course is
A.1.7	I think the course description's prerequisites are
A.1.8	In general, I think this is a good course
A.2.1	I think my English skills are sufficient to benefit from this course

Table 3.2: Questions in Form B

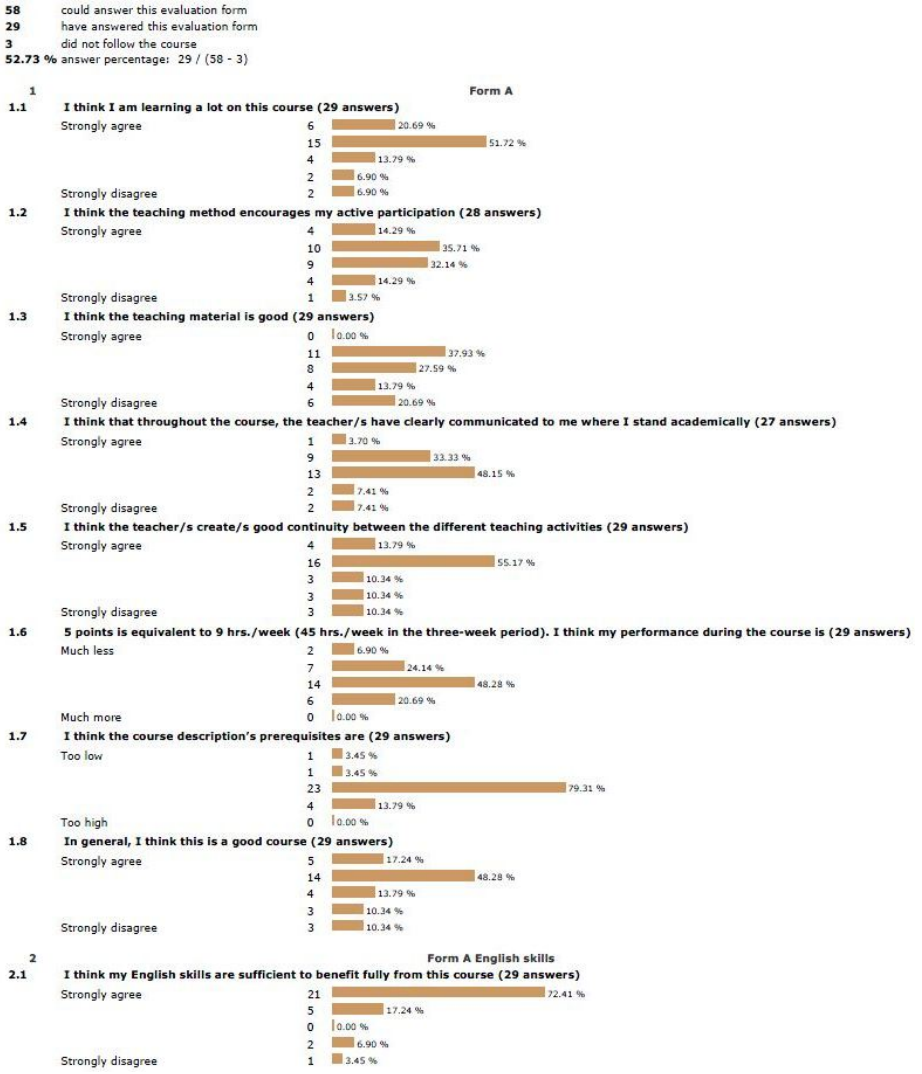
	Question
B.1.1	I think that the teaching gives me a good grasp of the academic content of the course
B.1.2	I think the teacher is good at communicating the subject
B.1.3	I think the teacher motivates us to actively follow the class
B.2.1	I think that I generally understand what I am to do in our practical assignments/lab courses/group computation/group work/project work
B.2.2	I think the teacher is good at helping me understand the academic content
B.2.3	I think the teacher gives me useful feedback on my work

Additionally, the course responsible has the possibility of adding some extra questions to the standard course evaluation survey. It is not mandatory to fill out the course evaluation at DTU, but students get a couple of reminders about the evaluations during the last week of the semester.

The results of Forms A and B are summarized as histograms showing the distribution of answers for each question (Figure 3.1). The lecturer can see the evaluation response rate, the histograms of the answers for the course evaluation, his own teaching evaluation, evaluation of the teaching assistants of the course and student answers to open-ended questions. It is up to the lecturer to share the results of the evaluation with the students.

The results of the course evaluations for the last 5 years are also available for

Figure 3.1: Example of results of course evaluation



the students on DTU's course catalogue. It can be used by potential students when they decide which courses to choose and when.

3.2 Analysis of the quantitative data from student evaluations

The results of course evaluations, that are summarized in histograms are later used by the teachers to improve their teaching, by future students while choosing courses, and by the department's study board for personnel decisions. However, it is obvious that the evaluations contain much more information than can be extracted just from the histograms.

As the first stage of the research, the degree of association between the two quantitative parts of the DTU's evaluation survey: the evaluation of the course and the evaluation of the teacher, was investigated.

The data from the evaluation of the Introduction to Statistics course in year 2007 and 2008 was used in the investigation. Descriptive statistics of the data are presented in the section 7.3.1. For the analysis of consistency of the results over time, the evaluation data for years 2009-2012 has been analyzed in the thesis. The second part of the form B is not used for the evaluation of the teacher of this course. It is common practice at DTU to have only the first 3 questions for evaluation of the main lecturer for large courses, while questions B.2.1 - B.2.3 are used for teaching assistants. Therefore, another course was analyzed at the later stage of the research.

One of the largest courses at DTU where the teacher is evaluated using all 6 questions from form B is the Introductory Programming with Matlab course. The course is available 4 times per year: twice as a 13-week course (fall and spring semesters) and twice as an intensive 3-weeks course (January and June). In chapter 8 the 3-week course held January 2010 was used for analysis. Descriptive statistics of the data are presented in section 8.5.1. The number of students that follow the course is very different from semester to semester. Here we will focus on the intensive 3-week version of the course. June courses are more popular (approximately 300 students) than the January courses (around 100-150 students). Figure 12.1 shows the number of students registered for the course and the course evaluation response rate and figure 8.3 presents the average SET scores of the evaluation of the course (Form A) over the period from January 2010 to June 2013.

The response rates ranges form 19% to 55%. Therefore, it is impossible to check the consistency of correlation structures every year due to lack of observations in some years. For the comparison of methods we use results from one semester (January 2010), and for the robustness study we examine the same course at two other time points (June 2011 and June 2012).

3.3 The Mid-term Experiment

In order to check whether the midterm evaluations can lead to improvement within the semester to meet the needs of the students in a specific class, and not just future students, an additional mid-term course and teacher evaluation was set up for the largest courses at DTU.

In order to conduct the experiment around 40 courses were needed, that would be divided in to two groups. In 20 courses the course teacher would be allowed to see (and act upon) the midterm evaluation, while in another 20 courses the results of the midterm evaluation would be kept secret until the end of the semester.

The heads of main DTU Departments and the heads of the Department study boards were contacted to provide courses which could and would be willing to participate and that satisfied the following criteria:

1. The expected number of students for the course should be more than 50.
2. There should be only one main teacher in the course.
3. The course should not be subject to other teaching and learning interventions (which often imply additional evaluations).

Unfortunately, not all the teachers of courses that satisfied the criteria, were willing to participate in the experiment. A study was conducted during the fall semester of 2010 and included 35 courses. The majority of the courses were introductory bachelor level courses, but also a few master's courses were included. The courses were taken from six different Departments: Chemistry, Mechanics, Electronics, Mathematics, Physics, and Informatics. The list of courses under experiment is provided in Table 3.3 (courses where the teacher got the results of the mid-term evaluation) and Table 3.4 (courses where the teacher did not get the results of the mid-term evaluation)

An extra midterm evaluation, identical to the end-of-term evaluation, was set up for all the selected courses in the 6th week of the semester. The end-of-term evaluations were conducted as usual in the 13th week of the semester.

The courses were randomly split into two groups: one half where the teacher had access to the results of the midterm evaluations (both ratings and qualitative answers to open questions) and another half where that was not the case (the control group). The courses were split such that equal proportions of courses

Table 3.3: List of DTU courses under mid-term Experiment where teachers had an access to midterm evaluations

Course	# of resp. mid - final (matches)	Course response rate mid - final	Matches as % of all matches	Access to midterm evaluations
Organic Chemistry 2	51 - 52 (41)	49% - 50%	3.1%	Yes
Statics	62 - 37 (32)	56% - 34%	2.48%	Yes
Plate and Shell Structures	26 - 25 (20)	58% - 58%	1.55%	Yes
Fracture Mechanics	31 - 28 (27)	65% - 60%	2.09%	Yes
Electric Circuits 1	54 - 45 (34)	46% - 39%	2.64%	Yes
Advanced Engineering Mathematics 2	243 - 212 (167)	45% - 41%	12.96%	Yes
Geometric Operations in Plane and Space	49 - 36 (29)	56% - 41%	2.25%	Yes
Physics 1 (class a)	87 - 86 (75)	29% - 29%	5.82%	Yes
Introduction to Statistics	120 - 158 (101)	38% - 51%	7.84%	Yes
Development Methods for IT-Systems	29 - 15 (10)	32% - 17%	0.76%	Yes
Probability Theory	27 - 22 (17)	28% - 24%	1.32%	Yes
Introductory Programming	29 - 15 (9)	28% - 13%	0.70%	Yes
Data Security	39 - 29 (27)	44% - 33%	2.09%	Yes
Optimization and Data Fitting	43 - 39 (29)	56% - 53%	2.25%	Yes
Multivariate Statistics	42 - 35 (29)	53% - 47%	2.25%	Yes
Programming in C++	37 - 25 (22)	49% - 34%	1.71%	Yes
Web 2.0 and mobile interaction	39 - 25 (18)	43% - 28%	1.40%	Yes
TOTAL (17 courses)	687	-	53.3%	Yes

Table 3.4: List of DTU courses under mid-term Experiment teachers did not had an access to midterm evaluations

Course	# of resp. mid - final (matches)	Course response rate mid - final	Matches as % of all matches	Access to midterm evaluations
Inorganic Chemistry	23 - 23 (20)	40% - 43%	1.55%	No
Physical Chemistry 2	38 - 29 (28)	57% - 45%	2.17%	No
Hydrodynamics	28 - 31 (26)	49% - 56%	2.02%	No
Computational Fluid Dynamics	37 - 30 (28)	74% - 61%	2.17%	No
Mechanics	47 - 28 (22)	50% - 30%	1.71%	No
Electronics	22 - 12 (8)	30% - 16%	0.62%	No
Engineering Electromagnetics	29 - 36 (24)	48% - 60%	1.86%	No
Calculus and Algebra 1	216 - 146 (111)	47% - 32%	8.61%	No
Calculus and Algebra 2	70 - 62 (42)	33% - 29%	3.26%	No
Physics 1 (class b)	23 - 25 (20)	32% - 35%	1.55%	No
Physics 1 (class c)	101 - 81 (72)	47% - 38%	5.59%	No
Probability and Statistics	69 - 83 (52)	29% - 36%	4.03%	No
Introductory Programming with Matlab	69 - 79 (52)	42% - 48%	4.03%	No
Introduction to Numerical Algorithms	23 - 18 (14)	37% - 33%	1.09%	No
Windows Programming using C# and and .Net	35 - 29 (14)	43% - 36%	1.09%	No
Digital Electronics 1	42 - 33 (26)	53% - 42%	2.02%	No
Software Development of Web Services	31 - 21 (13)	42% - 30%	1.01%	No
Embedded Systems	44 - 37 (30)	69% - 51%	2.32%	No
TOTAL (18 courses)	602	-	46.7%	No

within each Department were assigned to the two groups. The students of all 35 courses were also informed that only half of the involved courses/teachers would get to see the mid-term evaluation immediately after the evaluation.

The results of the mid-term evaluation were kept secret for everyone, except teachers of the selected courses. These teachers got the printed versions of the mid-term evaluation results.

In order to be able to couple student responses in mid-term and final term and keep students anonymous, a static encryption key was developed by CampusNet IT support.

3.4 Students Written Feedback

It is obvious that student answers to the open ended questions add extra information about course weaknesses and successes. In order to utilize this additional information from student comments some text mining methods should be applied.

At the Technical University of Denmark, in addition to quantitative 5-point Likert scale questions (Table 3.1 and Table 3.2) there are also 3 qualitative questions where students can type their feedback (Form C):

C.1.1 What went well?

C.1.2 What did not go so well?

C.1.3 Suggestions for changes?

After the evaluation period is finished, the teacher of a course can see all the students comments, as a text file, together with histograms of the quantitative results of the evaluation.

To find a proper course to analyze, the results of the evaluations of the largest courses taught at DTU were checked for the number of comments and length of comments. The period of investigation was from fall semester 2007 until spring semester 2012.

As a result, the Mathematics for Engineers II course was selected for analysis. The course is a bachelor level taught by Department of Mathematics and is a 5-ECTS points introductory level course that is available in both spring and

fall semesters. The course is well established, the curriculum, the book and the structure of the course is the same over the period. The number of students that followed the course during spring semesters is approximatively 250 and during fall semesters is approximately 500. The course is mandatory for all the students who want to enter a Masters program at DTU. According to the program the most convenient is to take this course in the fall semester of the second year of education. A part of the spring semester students are those who failed the course in the fall semester.

Figure 3.2: Response rates for quantitative and quantitative part of evaluation

Figure 3.2 presents the response rates on the course for the period of investigation, from fall 2007 to spring 2012. The response rates are lower for spring semesters (33-49%), than for fall semesters (41-62%). Response rates on the open-ended questions is much lower than on the quantitative questions. There are more students who write positive comments than those who write negative. However the average length of the negative comments (35 words) is 10 words larger than the average length of positive comments (26 words) and suggestions (25 words). Comment length varies from a single word to a paragraph with detailed analysis of approximately 150 words.

The red line A.1.8 on Figure 3.2 represents the average score the course get on the question about overall satisfaction with the course. The student satisfaction

of the course dropped down by approximately half a point on a Likert scale (Likert, 1932) in spring 2011. This is mainly due to the fact that one of the main teachers changed in spring 2011. This caused a drop in course rating, since the teacher was not experienced in teaching introductory-level courses and had higher expectations to the students. The results of course and teacher evaluations were analyzed and changes in teaching style were made for the next semesters.

The general objective of the course is to provide participants with tools to solve differential equations and systems of differential equations. The course content includes: solution of homogeneous/inhomogeneous differential equations and systems of differential equations, transfer functions, infinite series, power series, Fourier series, applications of infinite series for solving differential equations, the exponential matrix, stability and introduction to nonlinear differential equations. Students also learn how to use Maple to solve problems on the above topics. Some of the above mentioned mathematical issues are mentioned in students comments.

3.5 Student demographic data

As was mentioned before, student evaluations at the Technical University of Denmark are performed using web-based forms via DTU's intranet - CampusNet. One of the advantages of on-line based evaluation of teaching is that the results can be combined with demographical data while keeping the students' anonymity as well as with course specific characteristics.

DTUs intranet is developed by a third party - a private company named Arcanic A/S, that is located at the DTU campus. Arcanic's core competence is to develop and implement education-specific IT systems. The company is also responsible for DTU's student and course databases.

In order to combine the results of evaluation with student-specific characteristics and keep student anonymity, the same static encryption key was used as for the mid-term experiment.

Among course characteristics: course size, experience of the teacher with the course, ECTS points, course level and course language were available. Among student specific characteristics that are available at DTU are age, gender, nationality, study program, study line, weather the student takes the course for the first time and obtained grade. Table 12.3 and table 12.2 present the course-specific and student-specific characteristics for the students and courses, that

participated in the mid-term evaluation experiment.

The administration at DTU also has a dataset with information about students who studies at the university. The dataset provides: year of entering the university, type of entrance exam, high school, high school GPA, university GPA, grades on mathematics, physics, and chemistry from high school.

Chapter 4

Methods

Analysis of the results of student evaluation has become quite popular during the last 10-15 years. Since many universities switched from "old-style" paper based evaluation form to more modern online-based evaluation forms obtaining and processing the student ratings data has become easier. Depending on the structure of the data and question of interest various type of statistical methods can be applied, starting from simple ANOVA tables or Student's t-test to complicated behavioural types of analyses used in social science.

This chapter briefly describes statistical methods used during the project. Table 4.1 presents a list of methods used together with the papers they were used in. The table contains statistical methods for analysing quantitative data as well as some text-mining tools used for analysis of student comments.

4.1 Statistical methods of analysing the quantitative results of course evaluation

Different statistical methods are applied to analyse the quantitative data from the students evaluation of course and teaching quality. Summaries and calculations are often made by teachers, who want to better understand their own eval-

Table 4.1: List of methods used

Section	Method	Chapter
4.1.1	Student's t-test	Chapter 11
4.1.3	Factor Analysis	Chapter 10
4.1.4	Logistic regression	Chapter 10
4.1.5	Canonical Correlation Analysis	Chapter 7, 8
4.1.6	Regularized Canonical Correlation Analysis	Chapter 8
4.1.7	Sparse Canonical Correlation Analysis	Chapter 8
4.1.8	Cross-validation	Chapter 8
4.2.3	k-means	Chapter 9
4.2.4	Latent Semantic Indexing	Chapter 9
4.2.5	Key Phrase Extraction	Chapter 10
4.2.6	Stemming	Chapter 9, 10

uations or test their own hypotheses. The most common methods are analysis of variance (ANOVA), Student's t-tests, ordinary least squares (OLS) regression, correlation analysis, principal components analysis (PCA), factor analysis, logistic regression, mixed models and others.

4.1.1 Student's t-test

Student's t-test is any statistical hypothesis test in which the test statistic follows a Student's t distribution if the null hypothesis is supported. It is commonly used to determine if two sets of data are significantly different from each other in mean. Among the most frequently used t-tests are:

- A one-sample location test of whether the mean of a normally distributed population has a value specified in a null hypothesis (one-sample t-test).
- A two-sample location test of the null hypothesis that the means of two normally distributed populations are equal (two-sample t-test).
- A test of the null hypothesis that the difference between two responses measured on the same statistical unit has a mean value of zero (paired t-test).
- A test of whether the slope of a regression line differs significantly from zero.

Two-sample t-tests for a difference in mean involve independent samples, paired samples and overlapping samples. Paired t-tests are a form of blocking, and

usually have greater power than unpaired tests (Zimmerman, 1997).

4.1.2 Principal Component Analysis

Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. It was developed by Pearson (1901) and Hotelling (1933).

PCA is a linear transformation that transforms the data to a new coordinate system such that the new set of variables, the principal components, are linear functions of the original variables. This transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the previous components. The number of principal components is less than or equal to the number of original variables.

PCA is closely related to factor analysis and to canonical correlation analysis.

4.1.3 Factor analysis

Multivariate data often include a large number of measured variables, and often those variables "overlap" in the sense that groups of them may be dependent. In statistics, factor analysis is one of the most popular methods used to uncover the latent structure of a set of variables. This method helps to reduce the attribute space from a large number of variables to a smaller number of unobserved (latent) factors. Factor analysis was developed in the field of psychology, but it is also applied in many other fields (Vincent, 1953).

The most popular form of factor analysis is exploratory factor analysis (EFA), that is used to uncover the underlying structure of a relatively large set of variables. The researcher's a priori assumption is that any indicator may be associated with any factor.

Factor analysis searches for joint variations in response to unobserved latent variables. The observed variables are modeled as a linear combinations of the potential factors, plus an "error" term. The coefficients in a linear combination are called factor loadings. The information gained about the dependencies between observed variables can be used later to reduce the set of variables in a

dataset for further analysis.

Sometimes, the estimated factor loadings can give a large weight on several factors for some of the observed variables, making it difficult to interpret what those factors represent. The varimax rotation is the most commonly used criterion for orthogonal rotation, that helps to simplify the structure and ease interpretation of the resulting factors (Hair et al., 2006).

4.1.4 Logistic regression

Logistic regression is a type of regression analysis used in statistics for predicting the outcome of a categorical dependent variable based on one or more usually continuous predictor variables. In cases where the dependent variable consists of more than two categories which can be ordered in a meaningful way, ordered logistic regression should be used.

The relationship between a categorical dependent variable and independent variables is measured, by converting the dependent variable to probability scores. The model only applies to the data that meet the proportional odds assumption, that the relationship between any two pairs of outcome groups is statistically the same. The model cannot be consistently estimated using ordinary least squares; it is usually estimated using maximum likelihood (Greene, 2006).

4.1.5 Canonical Correlation Analysis

Canonical correlation analysis (CCA), introduced by Hotelling (1935, 1936), is a common method used to investigate the degree of association between two sets of variables in a linear sense, and can also be used to produce a model equation which relates the two sets of variables.

The method considers two matrices X and Y of order $n \times p$ and $n \times q$ respectively. The columns of X and Y correspond to variables and the rows correspond to experimental units. CCA assumes $p \leq n$ and $q \leq n$, and that matrices X and Y are of full column rank p and q, respectively. The main idea behind CCA is to find canonical variables in the form of two linear combinations:

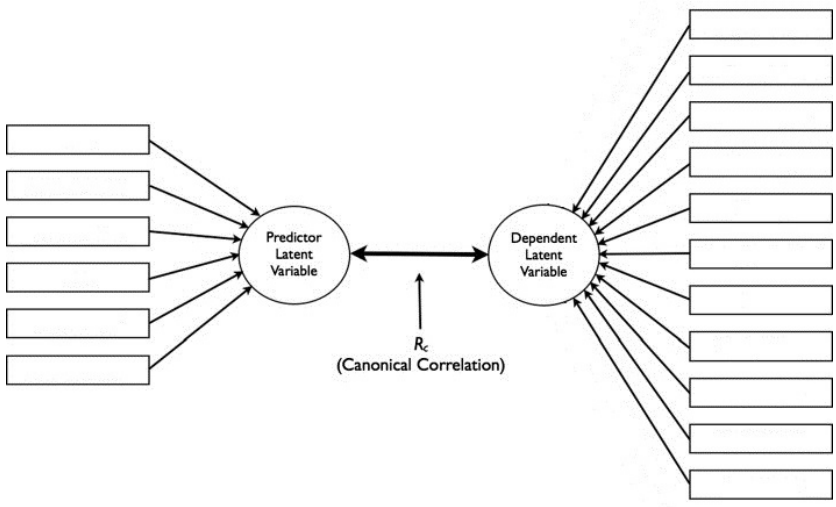
$$\begin{aligned} w_1 &= a_{11}x_1 + a_{21}x_2 + ... + a_{p1}x_p \\ v_1 &= b_{11}y_1 + b_{21}y_2 + ... + b_{q1}y_q \end{aligned} \tag{4.1}$$

such that the coefficients a_{i1} and b_{i1} maximize the correlation between two canonical variables w_1, and v_1. In other words, the problem consists in solving

$$\rho_1 = cor\,(v_1, w_1) = \max_{a,b} cor\left(a^T X, b^T Y\right) \tag{4.2}$$

Figure 4.1 provides the illustration of canonical correlation method and section 8.3.1 provides more details regarding the method.

Figure 4.1: CCA illustration



4.1.6 Regularized Canonical Correlation Analysis

CCA cannot be performed when variables $x_1, x_2, ..., x_p$ and/or $y_1, y_2, ..., y_q$ are highly correlated. In this case the correlation matrices, that are used in the computational process, tend to be ill-conditioned and their inverses unreliable. To deal with this problem a regularization step can be included in the calculations.

The principle of ridge regression, developed by Hoerl and Kennard (1970), which shrinks the weights by imposing a penalty on their size can be incorporated to the CCA settings. In order to choose "good" values of regularization parameters a standard K-fold cross-validation procedure 4.1.8 can be used. More details of the methods are described in section 8.3.2.

4.1.7 Sparse Canonical Correlation Analysis

Sparse CCA (SCCA) is an extension of CCA that addresses another weaknesses of the classical CCA method. CCA cannot be performed when the number of observations is less than the greatest number of variables in both data sets ($n < max(p; q)$). In such case, a selection of variables should be performed jointly with the analysis of the two data sets. SCCA can also help to solve the problem of interpretability providing sparse sets of associated variables. These results are expected to be more robust and generalize better outside the particular study.

New, recently developed penalization methods such as lasso, developed by Tibshirani (1996), and elastic net, developed by Zou and Hastie (2005), can be incorporated to canonical correlation analysis. Lasso is a penalization method that shrinks coefficients to zero. Similarly to regularized CCA, where ridge regularization helps to solve instability due to multicollinearity, it is possible to introduce Lasso, which selects variables by putting weights to zero, into CCA 8.3.3.

4.1.8 Cross-validation

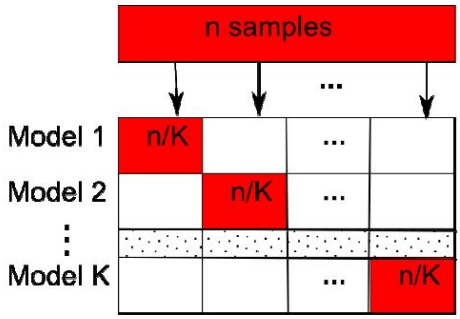
Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. The idea of the method is to divide the data into two segments: one used to learn or train a model and the other used to validate the model (Hastie et al., 2001).

Common types of cross-validation are:

- k-fold cross-validation. The original sample is randomly partitioned into k equal (or nearly equal) size subsamples (folds). Of the k subsamples, a single subsample is retained as validation data for testing the model, and the remaining $(k-1)$ subsamples are used as training data (Figure 4.2). Subsequently k iterations of training and validation are performed such that within each iteration a different fold of the data is held-out for validation while the remaining folds are used for learning.

- Leave-one-out cross-validation, that involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data.

Figure 4.2: K-fold cross-validation



Cross-validation can be applied in three contexts: performance estimation, model selection, and tuning learning model parameters.

4.2 Text-mining methods

Text mining is a process of deriving high-quality information from unstructured text. Text mining is an interdisciplinary field that draws on information retrieval, data mining, machine learning, statistics, and computational linguistics.

Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output. Typical text mining tasks are text clustering, concept extraction, production of granular taxonomies, sentiment analysis, document summarization and others.

4.2.1 Text pre-processing

Text pre-processing is the task of converting a raw text file into a well-defined sequence of linguistically-meaningful units. Text pre-processing is an essential part of natural language processing, since the characters, words, and sentences identified at this stage are the fundamental units passed on to further text-mining stages (Palmer, 2010).

Preprocessing text, that is also known as called tokenization or text normalization, includes some or all of the following steps:

- Throw away unwanted elements (e.g., HTML tags, (sometimes) numbers, UUencoding, etc.)

- Define word boundaries: white space and punctuations – but words like Ph.D., isn't, e-mail are problematic.

- Stemming (Lemmatization): This is optional. Often (but not always) it is beneficial to map all inflected word forms into the corresponding stem (4.2.6).

- Stopword removal: the idea is to remove words that occur in "all documents'. The most frequent words often do not carry much information. For example words "the", "a", "of", "for", "in", etc. The concept was first introduced b in 1958 by Luhn (1958).

4.2.2 Term-document matrix

A lot of text mining methods are based on construction of a term-document matrix, a high-dimensional and sparse mathematical matrix that describes the frequencies of terms that occur in a collection of documents. There are various ways to determine the value that each entry in the matrix should take.

Term frequency - inverse document frequency (tf-idf), is a numerical value which reflects importance of a word for a document in a collection of documents. The tf-idf value increases proportionally to the number of times a word appears in the document, but with an offset by the frequency of the word in the corpus. This helps to control for the fact that some words are generally more common than others (Salton and Buckley, 1988).

Tf-idf is defined as the product of two statistics: term frequency (the number of times that term occurs in a document divided by the total number of words in the document), and inverse document frequency (a measure of whether the term is common or rare across all documents). It is defined by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that ratio.

The tf-idf weight of term t in document d is highest when t occurs many times within a small number of documents, lower when the term occurs fewer times in a document, or occurs in many documents, and lowest when the term occurs in almost all documents of a collection.

4.2.3 Text clustering

Text clustering is one of the fundamental functions in text mining, used to divide a collection of texts into groups (clusters) so that documents in the same category group describe the same topic. Clustering text data faces a number of challenges like: the volume of text data, the dimensionality of data, sparsity and complex semantics.

The k-means clustering algorithm, developed by MacQueen (1967), is known to be efficient in clustering large data sets. It is also one of the simplest and the best known unsupervised learning algorithms. K-means clustering helps to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. The best number of clusters k is not known *apriori* and must be computed from the data. Usually, as the result of a k-means clustering analysis, means for each cluster on each dimension (usually called centroids) are analyzed.

There are various modification of k-means algorithms. Another algorithms often used in text clustering are various hierarchical clustering methods that are more accurate, but usually suffers from efficiency problems.

4.2.4 Latent Semantic Indexing

Latent Semantic Indexing (LSI) is an indexing and retrieval method that uses a mathematical technique called singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. Search engines, Internet Marketing Professionals, and Website Designers often use LSI in their day-to-day activities.

LSI is based on the principle that words that are used in the same contexts tend to have similar meanings. A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts (Deerwester, 1988).

LSI begins by constructing a term-document matrix, to identify the occurrences of the unique terms within a collection of documents. In a term-document matrix, each term is represented by a row, and each document is represented by a column, with each matrix cell, initially representing the number of times the associated term appears in the indicated document. This matrix is usually very large and very sparse. Then a rank-reduced SVD is performed on the matrix to determine patterns in the relationships between the terms and concepts con-

tained in the text. Efficient LSI algorithms only compute the first k singular values and term and document vectors as opposed to computing a full SVD and then truncating it.

4.2.5 Keyphrase extraction

Extraction of keyphrases is a natural language processing task for collecting the most meaningful words and phrases from the document. It helps to summarize the content of a document in a list of terms and phrases and thus provides a quick way to find out what the document is about. Automatic keyphrase extraction can be used as a ground for other more sophisticated text-mining methods.

There are various types of methods, that help to extract key- words and phrases from the text. In this study, a simple language-independent method, called Likey (Paukkeri and Honkela, 2010), is used. The only language-specific component is a reference corpora. According to the method, a *Likey ratio* (10.1) is assigned to each phrase (n-gramm)

$$L(p,d) = \frac{rank_d(p)}{rank_r(p)} \tag{4.3}$$

where $rank_d(p)$ is the rank value of phrase p in document d and $rank_r(p)$ is the rank value of phrase p in the reference corpus. The rank values are calculated according to the frequencies of words of the same length n. The ratios are sorted in increasing order and the phrases with the lowest ratios are selected. Phrases occurring only once in the document cannot be selected as keyphrases.

4.2.6 Stemming

Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form. The stem is not necessarily identical to the morphological root of the word. For example, stemming reduces the words "fishing", "fished", "fish", and "fisher" to the root word, "fish". On the other hand, "argue", "argued", "argues", "arguing", and "argus" reduce to the stem "argu", but "argument" and "arguments" reduce to the stem "argument".

The first published stemmer was created by Lovins in 1968. Later a more popular stemmer was written by Porter (1980). The Porter stemmer is very widely

used and became the de facto standard algorithm used for English stemming. Porter released an official free-software implementation of the algorithm around the year 2000. Porter extended this work over the next few years by building Snowball, a framework for writing stemming algorithms, and implemented an improved English stemmer together with stemmers for several other languages.

Chapter 5

Results

This section briefly summarizes the results of the different aspects of student course evaluation that were investigated during the PhD project.

The project was started with the investigation of the degree of association between the two quantitative parts of evaluation survey, evaluation of the teacher and evaluation of the course (section 5.1.1). The next step was to address some of the weaknesses of the statistical method used in the first stage, namely use statistical tools that might produce more stable and generalizable results (section 5.1.2). Then, it was decided to conduct additional course and teacher evaluation in the middle of the semester. This was done in order to check whether the mid-term evaluations can lead to improvement within the semester to meet the needs of the students in a specific class, and not just future students (section 5.3). In parallel to conducting the mid-term experiment and analysing its results, different text-mining methods were applied in order to utilize the information from the students open-ended feedback and to combine it with the information obtained from the quantitative parts of the evaluation survey (section 5.2). Finally, students non-participation in evaluation surveys at DTU was investigated together with the relationships between students-specific characteristics and SET scores (section 5.4).

5.1 Association between student evaluations of courses and instructors

5.1.1 Canonical Correlation Analysis

As the first stage of the research project on student evaluation at the Technical University of Denmark, the degree of association between the two quantitative parts of DTU's evaluation survey was investigated. In the first part of evaluation (Form A) students evaluate different aspects of the course (Table 3.1), while on the second part (Form B) they evaluate the teacher (Table 3.2). Canonical correlation analysis (CCA) was performed on two samples from the Introduction to Statistics course namely in year 2007 and 2008. The detailed results are presented in chapter 7.

It was found that course and teacher evaluations are correlated. However, the structure of the canonical correlation is subject to some changes with changes in teaching methods from one year to another. The structures of the canonical correlations are presented in figure 5.1 and figure 5.2 for class 2007 and class 2008 respectively.

Figure 5.1: The structure of canonical correlation between the two parts of course evaluation in 2007

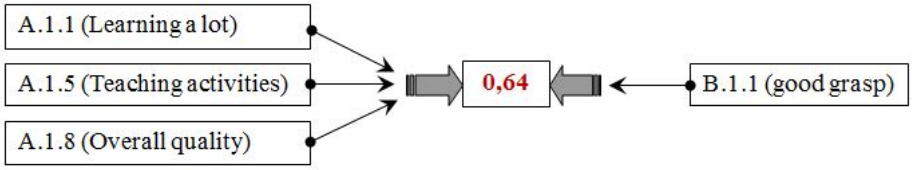
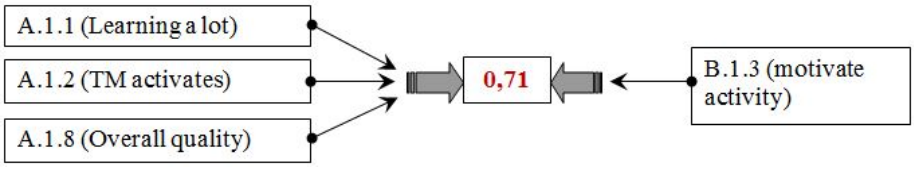


Figure 5.2: The structure of canonical correlation between the two parts of course evaluation in 2008



The presented structures are based on the canonical weights each variable (question of the survey) contributes to the unobserved latent variable, as well as on

canonical factor loadings and cross-loadings. More detailed information on how to interpret the results of canonical correlation analysis is presented in chapter 7 section 7.3.2.

The difference in structure of correlations can be explained by the change in teaching method from normal lectures in 2007 to combined lectures and video sequences, which could be replayed by the students in 2008. The video lectures were highly appreciated. Many students mentioned the video-recordings in their positive verbal feedback in form C.

One of the key elements of the reliability of SET is stability, which is the agreement between rates over time. The literature suggests that ratings of the same instructor tend to be similar over time (Overall and Marsh, 1980; Braskamp and Ory, 1994). Therefore, when the evaluation data for subsequent years became available, the CCA analysis was replicated for the course evaluation of the same course in 2009-2012. Figures 5.3 - 5.6 show the correlation structures between the two parts of evaluation surveys for each year.

Figure 5.3: Structure of the canonical correlation between the two parts of the course evaluation in 2009

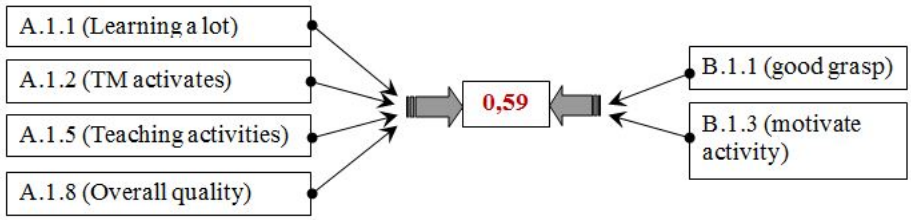
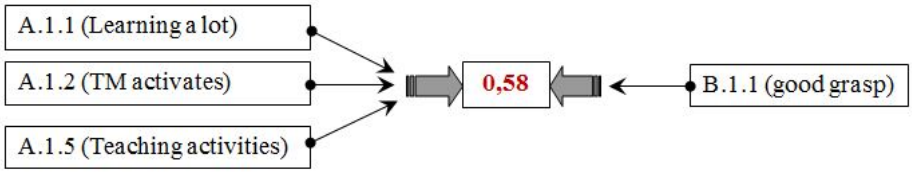


Figure 5.4: Structure of the canonical correlation between the two parts of the course evaluation in 2010



The figures suggest that the structure of correlation is relatively stable over time with slight changes from year to year. An overall conclusion that can be made is that the correlation between the students evaluations of the course

Figure 5.5: Structure of the canonical correlation between the two parts of the course evaluation in 2011

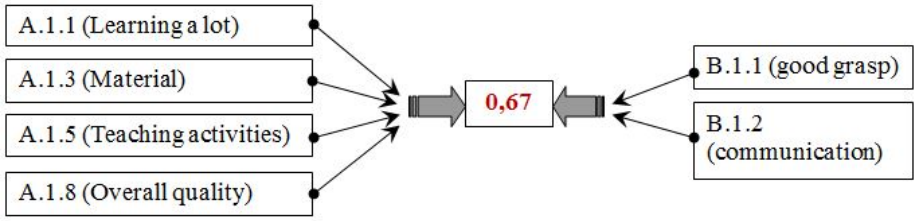
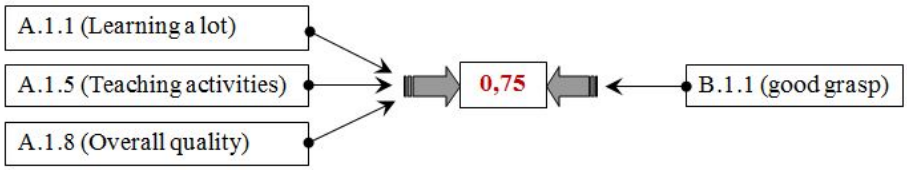


Figure 5.6: Structure of the canonical correlation between the two parts of the course evaluation in 2012



and of the teacher at the Introduction to Statistics course is mainly due to the relationship between the good continuity between teaching activities in the course(A.1.5), good content of the course (A.1.1) and good overall quality of the course (A.1.8) from one side and the teachers ability to give a good grasp of the academic content of the course (B.1.1) on the other side.

The changes might be due to the fact that even though course is one of the best rated courses at DTU, the instructor of the course continues to improve it every year. For example, in 2012, the Introduction to Statistics course was the first English-teaching bachelor level course at DTU. In addition to that, there was organized a live broadcasting of the lectures, so that students had a possibility to follow the lecture from any place with Internet connection.

The study has some weaknesses, that were fully or partially addressed in later investigations.

First of all, the second part of form B, namely questions B.2.1 - B.2.3 (Table 3.2) were not used for the course. It is common practice at DTU to have just the first 3 questions for the teacher evaluation for large courses. In such cases, the second part of the form B is active for the teaching

assistants only.

Second, there were high correlations between the scores of the questions within Form A and Form B, which might lead to wrong results.

Third, the CCA method does not take into account that the answers on student evaluations are categorical.

Fourth, the results are course specific and can not be generalized to other DTU courses.

5.1.2 Sparse and Regularized Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a common way to inspect the relationship between two sets of variables based on their correlation. However, the method produces inaccurate estimates of parameters, and non-generalizable results that are hard to interpret in case of insufficient sample sizes or high correlations between the variables in the data. Recently developed modifications of CCA, such as regularized CCA and sparse CCA, that impose L_2 and L_1 norm regularization respectively, were used to address such weaknesses. More details about the methods are presented in Chapter 8, sections 8.3.2 and 8.3.3.

All the three versions of CCA were used to analyze the the same sample of evaluation data for the Introductory Programming with MatLab course held on January 2010, the largest course, where all 6 questions of form B (evaluations of instructor) were active.

Figure 5.7 illustrates that the correlation between the students answers on different questions of evaluation survey were quite high.

The association between how students evaluate the course and how students evaluate the teacher was found to be quite strong in all three cases. However, regularized and sparse CCA produced results with increased interpretability over traditional CCA. The traditional CCA reported that the first 4 canonical corrections are statistically significant. This means that the structure of correlation lies in a 4 dimensional space, which is hard to visualize and interpret. The regularized and sparse CCA reported just one significant canonical correlation. The structures of these correlations presented in figure 5.8 and figure 5.9 respectively.

The two structures were similar, but the correlation structure resulting from the regularized CCA had more variables than the correlation structure for the

Figure 5.7: The correlations between the students answers on the student evaluation survey.

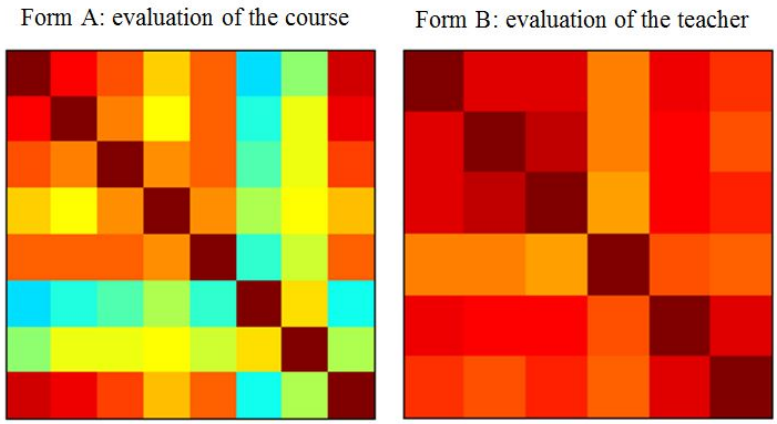


Figure 5.8: The structure of the regularized canonical correlation between the two parts of course evaluation.

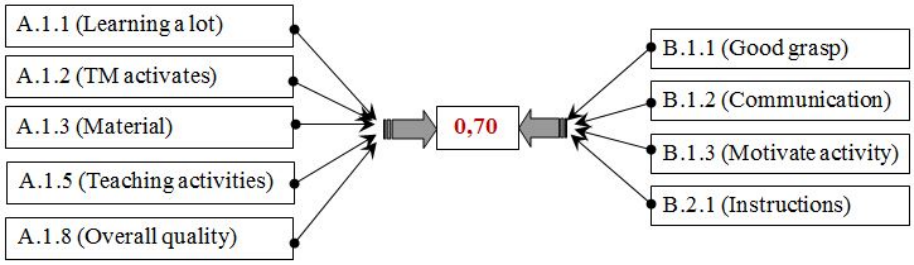
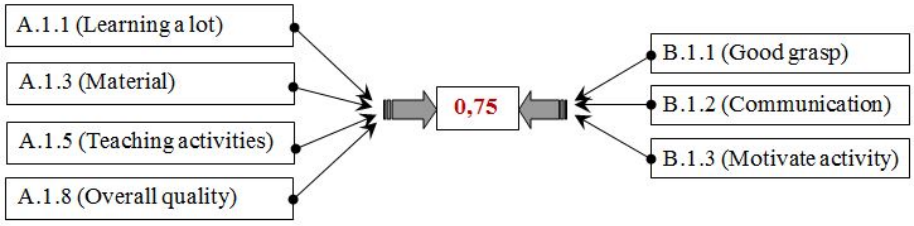


Figure 5.9: The structure of the sparse canonical correlation between the two parts of course evaluation.

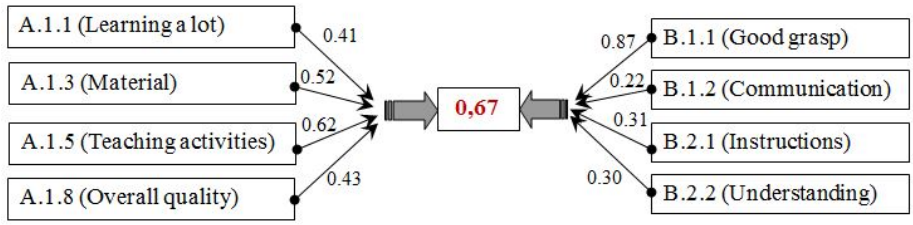


sparse CCA. This is mainly due to the fact that sparse CCA set the canonical weights of unimportant variables to zero, while the regularized CCA just shrinks these canonical weights, while the canonical factor loadings and cross-loadings can still show the importance of the variable.

The simplest model was obtained from the sparse canonical correlation analysis. The association between how students evaluate the course and how students evaluate the teacher was found to be due to the relationship between (on the course side, Form A) the good continuity between teaching activities in the course, content of the course, teaching material and overall quality of the course and (on the teacher side, Form B) the teachers ability to give a good grasp of the academic content of the course, teachers ability to motivate the students, and teachers good communication about the subject.

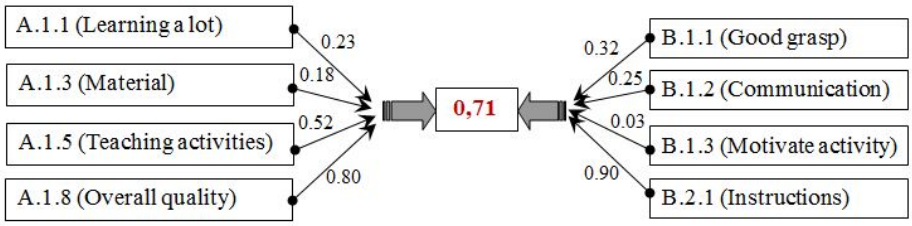
To check for the stability of the correlation structures, the subsequent years of the course should be analyzed. However, the introduction to programming with MatLab course had different numbers of students registered (from 100 to 350 students). The evaluation response rates range from 19% to 50%. Therefore, in some of the terms the number of observations was too small to conduct a proper analysis. Figures 8.7 and 8.8 present the correlation structures resulting from a sparse canonical correlation analysis of the evaluations in June 2011 and June 2012 respectively.

Figure 5.10: Structure of canonical correlation between the two parts of the course evaluation in June 2011



Overall, the two structures are similar. The only difference is on the side the evaluation of the teacher, where question B.2.2 (The teacher is good at helping me to understand the academic content) from the structure in 2011, while B.1.3 (The teacher motivates us to actively follow the class) was in the canonical correlation structure for 2012. The figures also show the weights each variable had in the latent canonical variable. The weights were different for the two years. However, it can be explained by the main teachers of the course being different.

Figure 5.11: Structure of canonical correlation between the two parts of the course evaluation in June 2012



The association between how students rate the teacher and the course was found to be subject to change with the change of teaching methods and with the change of teacher.

5.2 Text mining of student comments

The first part of the project considered the association between the two quantitative parts of the evaluation survey. However, many lecturers pointed out that the student written feedback provides more precise information of students points of satisfaction or dissatisfaction than the quantitative score. Moreover, student ratings of courses and teachers are subjective.

The current process of analysing SET results at DTU does not include analysis of students' open-ended feedback. The students' comments on what went well, what did not went so well and students' suggestions are available for course teachers and for university administration. The traditional way of analysing students' comments, i.e. reading, is hardly applicable when all courses of the university or department are analysed. In this situation, an automated method for extracting the most important information from the students written feedback may be able to provide insight to university administration and departments study boards on how a course was conducted, what went well, and what could be improved.

There are some challenges in analyzing the students written comments. First of all, the response rates on open-ended SET questions is usually below 20% (Jordan, 2011). Moreover, standard text-mining methods are developed for analysis of large documents or large collections of documents, while the students comments are different in length, ranging from just a few words to several

paragraphs of detailed discussion. Another challenge is that many comments contain slang, mistakes, misprints, word contractions, course-specific terms and abbreviations. However, one of the advantages of DTUs survey is that students write their positive and negative feedback separately.

5.2.1 Clustering students comments

Different text-mining methods have been tried during the project. As the first step, the analysis of positive student written feedback was done for the Introduction to Statistics course. This is the same that was analyzed in chapter 7 for the association between the students evaluation of the teacher and students evaluations of the course. The course is one of the best rated courses at DTU, therefore the number of negative comments is very small.

Chapter 9 presents analysis of the positive comments for two subsequent years, applying k-means clustering and singular value decomposition. It was found that changes in teaching methods was reflected in the students written feedback. In particular, comments about the introduction of video lectures formed a separate cluster of comments in 2008, while in 2007 the main topics of comments were about: overall quality of the teacher and the course, teachers ability to convey the subject, good quality and content of the lectures, and good teaching assistants.

A conclusion that can be made is that students react on changes in teaching methods. The study suggests that analysis of whether the teacher improves his/her course over the years can be done by analyzing the students written comments.

The study also demonstrates the limitations of using text-mining methods on student comments. The singular vectors obtained from SVD are usually used for further query matching, when the new texts are added to the collection of documents. However, due to changes the teacher makes from year to year, the the basis of singular vectors obtained from student comments in one semester may not be relevant for subsequent years. In order to build a good basis for query matching, a number of courses should be used, to be able to address different aspects of teaching.

5.2.2 Clustering courses based on students comments

The second step was to try to cluster different DTU courses based on the positive and negative comments. The idea was to find clusters, that could reveal courses with similar problems or similar successes. However, the results of k-means clustering were not satisfactory, since course specific terms like programming language name (Matlab, splus, SAS, etc) was found to be dominant words in centroids.

As the result, the obtained clusters mostly represented courses with similar topics. For example, one of the clusters was formed primarily by different kinds of chemistry courses, or courses that use the same tools to solve exercises, for example, courses that used Matlab or Maple.

Another idea was to try to cluster every single student comment. Students of the same course might have different complaints and different preferences, while one is not satisfied with a book, another might be not happy with the teaching assistant.

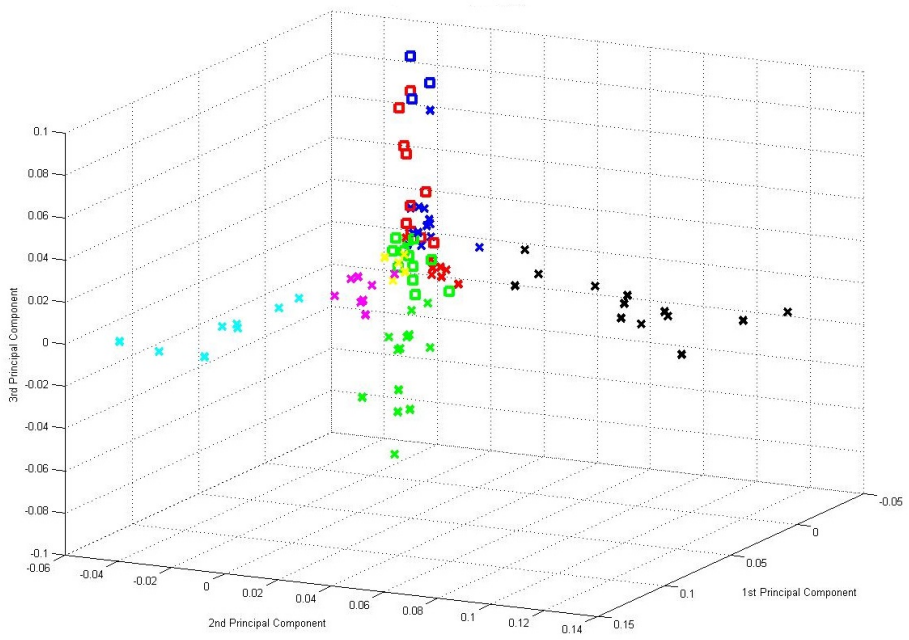
The k-means algorithm, unfortunately, does not guarantee that a global minimum in the objective function will be reached. This is a particular problem in document clustering if a collection of documents contains outliers. I.e. documents that are far from any other documents and therefore do not fit well into any cluster. Moreover, the number of clusters has to be defined in advance and the algorithm is dependent upon the starting centroid locations.

Figure 5.12 illustrates the problem. Centroids of some of the clusters where hard to interpret and the documents (comments) in those clusters where long and quite far from each other. While short one or two sentence comments formed quite reasonable clusters.

Another attempt was to use the k-nearest neighbors method, to find comments that are close to some important predetermined query, for example to the query "book is good". This approach helped finding comments regarding the course issue. However long comments with detailed analysis of different course issues where assigned to either multiple clusters or to no cluster at all.

Moreover, all these clustering methods do not help to build the relationships between the students written feedback and the quantitative score that student give to the course in the evaluation survey.

Figure 5.12: Clusters of students comments plotted on the first 3 principal components



5.2.3 Relationships between students written comments and quantitative SET scores

Clustering the students comments can provide an overview of the main topics, students address in their comments. However, it does not provide information about how the points of students satisfaction or dissatisfaction, that are addressed in the written comments, are related to the scores in quantitative part of the evaluation survey. For example, a student might not be satisfied with the quality of the textbook. He/she mentions in his/her feedback. The question is how it effect the score he/she gives to the quantitative question about teaching material (A.1.3). To do so the comments should be transformed into numbers, and then some kind of statistical model should be built.

In order to find the main topics of the students' comments, a keyphrase extraction tool was used. Based on the extracted key terms, the qualitative feedback was transformed into quantitative data. Since the number of key terms was quite high, factor analysis helped to reveal the underlying structure of the stu-

dents' written comments. Moreover, the method also helped to detect outliers. These are the comments that are very different from the others, like a negative comment in a collection of positive feedback.

Among the positive comments the extracted factors are: teacher qualities, weekly home assignments and its level, textbook quality, blackboard teaching, teaching assistants communication, weekly question sessions, overall course quality in comparison with other DTU courses. In the negative comments the most interesting factors are: gradings of home assignments, lack of examples during lectures, projects workload, Maple as a tool to solve exercises, English-speaking TAs, lack of TAs and course administration issues.

A regression analysis showed that some of the revealed factors, extracted from the positive and from the negative comments, had a significant impact on how students rated different aspects of the course. The relationship between the factors and students' overall satisfaction with the course (A.1.8) changes from semester to semester, meaning that teachers of the course take the comments into account and continue improving the course.

One of the improvements to this analysis, may be to incorporate students background into account. Students from the course under investigation have a very diverse background. The course is mandatory, for those students aiming at the master degree. Therefore, students with also poorer mathematical backgrounds have to follow it. Another improvement is to use more sophisticated text-mining method that takes the synonyms, like "teacher" and "lecturer", into account.

5.3 Effects of the mid-term SET on the end-of-semester SET

Student evaluations of teaching at the Technical University of Denmark have been performed online since 2001. The evaluations are collected during the last week of the semester before the final exam. Under this set up, the current students of the course do not benefit from the the end-of semester evaluations. Therefore, in the fall semester of 2010, an extra evaluation was set up for 35 selected courses on the 6th week of the course in addition to the end-of term evaluation. The evaluations contained the same questions as usually. Half of the teachers were allowed access to the midterm results. The study presented in chapter 11 analyzes the changes in SET scores from mid-term to end-of term evaluations.

The results illustrate that students are generally more satisfied with their courses

and teachers at end-of-term when midterm evaluations are performed during the course and teachers are informed about the results of the evaluations. Improvements related to student learning, student satisfaction, teaching activities, and communication showed statistically significant average differences of 0.1-0.2 points between the two groups. These differences are relatively large compared to the standard deviation of the scores when the student effect is removed (approximately 0.7). The same points of improvement for some courses are also reflected in written student comments, which illustrates the usefulness of midterm evaluations when addressing improvement of teaching for the current course students, and not just for future students.

The teachers were not obliged to take any specific actions based on the results from the mid-term evaluation. However, it turned out that almost $^3/_4$ of the teachers followed up on the evaluations by sharing the results with their students and/or making changes in the course for the remaining part of the semester. The major points of students dissatisfaction are reflected both at midterm and end-of-term comments. Criticism over the course book or the teaching assistants, can hardly be changed within semesters, but can more easily be changed from semester to semester.

Since the general experience is that response rates decrease when students are asked to fill in questionnaires more frequently, it seems to be preferable to conduct midterm evaluations as a standard questionnaire, instead of end-of-term. This is because the midterm evaluations capture both the points of students dissatisfaction that can be improved only for the future students, but also some weaknesses that can be addressed in the second part of the semester.

Another solution is to have both mid-term and end-of-term evaluations focusing on different aspects on the course. The midterm evaluation could focus on the formative aspect, questions concerning issues related to the teaching and learning process that can be changed during the semester, while the end-of-term evaluations could be reduced and focus on general questions and matters that are left out in the mid-term evaluation.

5.4 Non-participation in SETs

One of the problems of the course evaluations is that not all students, who participate in a course participate in the evaluation surveys. This non-participation can lead to biased results of the student evaluations. Chapter 12 presents the investigation on the non-response bias at DTU and together with the investigations of whether student and course characteristics have an impact on SET

scores.

The study considers the data from the mid-term experiment held at DTU in the fall 2010 (section 3.3). There was a little overlap in respondents at the two time points, but both time points showed response percentages close to "normal".

The results illustrate that different course specific and student specific variables have an effect on whether a student participates in the student evaluation of teaching in the middle or at the end of the semester. Some of the variables had similar effects at both end of term and midterm, while others had opposite directions, different magnitude or just no effect at one of the time points. Overall, the female students with high GPAs (from DTU and from the high school) who take the course for the first time were more likely to participate in both mid-term and end-of-term course evaluations.

The study also shows that the obtained grade was the factor that had a significant positive effect on both students ratings and students SET participation. Students with higher obtained grade were more likely to participate in course ratings. Moreover, students with higher grades tend to give higher rates to the course. However, the DTU GPA that had an influence on the SET participation appeared to have no impact SET scores.

Five variables: course department, course size, course weekday, students gender, and obtained grade, all had a significant impact on how students rate all or almost all the questions of the survey. Males tended to give lower ratings on evaluations than females in all the questions.

The findings concerning difference between mid-term and final term scores are similar to those found previously. The fact whether the teacher of the course knew the result of the mid-term evaluation had either no effect or negative effect on the ratings of the different course aspects.

Chapter 6

Discussion and Conclusions

In chapter 1 the objectives of the present thesis were presented, namely the application of statistical methods to analysis of student evaluations at the Technical University of Denmark. Questions that were addressed in this thesis include:

1. Is there a correlation and, if so, what is the structure of correlation between Form A (evaluation of course quality) and Form B (evaluation of the teacher)?
2. Which methods can be applied to find the most interpretable model of association between two quantitative parts of the evaluations (Form A and Form B)?
3. How can the information from student written feedback be utilized and what are the relationships between the information extracted from open-ended student comments and quantitative scores of SETs?
4. What is the effect of mid-term course evaluations on student satisfaction with the course as measured by end-of-term evaluations?
5. Which student specific characteristics and course specific characteristics can be the determinants of whether or not a student participate in the evaluation questionnaire? Are the students who submit an evaluation in the middle of the semester different from those who submit the evaluation at the end of the term?

6. Which student specific characteristics and course specific characteristics influences SET scores of courses?

This chapter summaries the scientific publications of this thesis, in Part II, together with the background and discussion of the research in Part I.

6.1 Discussion

The thesis investigated some of the number of issues of student evaluation of courses and teaching quality based on the students evaluation data from the Technical University of Denmark.

The findings include an investigation of the association between the two parts of the evaluation survey, the evaluation of the course and the evaluation of the teacher, investigation of the impact of the mid-term evaluation on the end-of term evaluation, and investigation of survey non-response bias at DTU. The work also considers application of text-mining methods to open-ended student feedback in order to find additional information that can be used for further deeper analysis of the evaluation results.

Student evaluation of the teaching process can be divided into three parts:

1. The process generally begins with an end-of-term evaluation questionnaire, where students are asked to rate the quality of the course and the teacher and/or provide qualitative feedback.
2. Next, summary statistics are produced and distributed to faculty and administrators.
3. Finally, the evaluation results are used by administration for personnel decisions, by faculty for improvement of the courses for future students and by potential students for course selection.

Each of these steps are highly debatable in the literature and are partially addressed in this thesis:

1. End-of-term evaluation is a general practice at DTU. However which questions should be asked and when is debatable.

- Low course evaluation response rates are often a concern for SET administrators. Shorter evaluation forms tend to have higher response rates and less missing values, than the long ones. This thesis investigated the association between evaluation of the course and evaluation of the teacher. The structure of this association was found to be relatively stable. SET administrators might consider to reduce the number of questions in the questionnaire in order to gain better response rates. However, that should be done very carefully. SETs must be multidimensional, in order to reflect multidimensionality of such a complex activity as teaching. According to Marsh and Roche (1997), the multidimensionality of SETs is based on the nine factors: Learning/Value, Instructor Enthusiasm, Organization/Clarity, Group Interaction, Individual Report, Breadth of Coverage, Examinations/Grading, Assignments/Readings, and Workload/Difficulty. The questionnaire, currently used at DTU is already small, but similar analysis could be used by other educational institutions.
- The thesis also considers the mid-term evaluation as an alternative to the end-of-term evaluation. The result shows that teachers, who had an access to the results of the mid-term evaluation managed to improve the courses for current students and get higher ratings at the end-of term evaluations. This finding is in line with findings by Cohen (1980), who concluded that on average the mid-term evaluations had made a modest but significant contribution to the improvement of teaching. There is evidence that student are more willing to participate in student evaluation, when they know that their opinion is valued. The mid-term evaluation shows the clear message to students that they can benefit in the second part of the semester by expressing their opinion in mid-term evaluations. Analysis shows that some of the course issues mentioned by students in the mid-term evaluation, can be improved in the second part of the semester, while others, like the textbook, can only be changed from semester to semester.

2. There are some issues regarding who get access to the evaluation results and when.

   - Faculty members whose course is being evaluated usually have full access to the collected data. At DTU it is up the teacher whether or not to share and/or discuss the evaluation results with the current students. Some teachers may use sharing of evaluation results with the students as an additional opportunity for collecting student feedback via in-class discussion of the SET results.
   - The university administration usually gets a summative overview as histogram plots or tables with average ratings, while the teachers have an opportunity to read the students written feedback. It is clear, that

for university study boards it is tedious to go through all comments for all courses. However, if there are enough comments, some information can be extracted from the students written feedback using text-mining methods. This information can provide additional insight to the university administration and department study boards on more precise points of students satisfaction and dissatisfaction, and probably answer the question "why?" SET scores are high or low for each course.

3. The results of evaluation should be properly analysed by both administration in personnel decisions and by faculty in course development decisions. Critics have argued that SETs alone may not improve teaching quality of an individual faculty member (Cohen, 1981).

   - Teachers can use the SET results for improvements of the course and for improvements of own teaching. Marsh (2007) indicated that students primarily evaluate the teacher rather than the course. The study found strong relationship between student evaluations of the courses and evaluations of the teachers. This can be a sign of the fact that better courses and therefore better SET results can be achieved in several different ways: improvement in a course can lead to better evaluation of the teacher, and improvement of the teacher qualities, can lead to better evaluation of the course.
   - Administration usually makes a comparison of the courses based on SET scores. However, as it was shown in this thesis as well as in other studies (Marsh, 2007; Cashin, 1995; Gravestock and Gregor-Greenleaf, 2008), certain course-specific characteristics may influence the ratings. The information about course characteristics, e.g., disciplinary field, class size, mandatory/elective, course level and workload should be considered when reviewing evaluation results. For example, elective courses tend to get slightly higher ratings than mandatory courses, especially if a mandatory course is outside a students major. All these effects, should be taken into account when comparing courses based on their evaluation scores.
   - Another issue is the quality of the data and amount of responses. Student non-response should be carefully studied. There is evidence that at DTU as well as at other universities (Kherfi, 2011; Feldman, 2007), high achievers have higher probability of participating in students ratings.
   - As a current practice, students submit the course evaluations at the end of the semester, which is one the busiest period of the whole semester. From this point of view it seems to be preferable to conduct mid-term evaluations, that also provides a valuable basis for adjustments, instead of end-of term evaluation. As an alternative

solution, mid-term evaluation questionnaires can be designed with some more focus on issues that can be changed during the semester, while an end-of-term questionnaire should capture the the overall full course impression. Both questionnaires would become more attractive towards the students if they contain a limited number of questions.

- The current DTU evaluation data collection processes can be improved via investigation of course drop-outs. Under the current set-up, there is no way to distinguish between the students who drop out from the class at the beginning of the course and who drop out right before the exam.

- Some literature suggests that student rating results should be considered in personnel decisions only when at least 10 students in a given class respond and only when the majority of the students in a class have completed the surveys. There are various methods to encourage students to participate in SET. The most popular are: sending students reminders about evaluations, offering extra credit, and spending class time filling out the survey.

For additional context, departments can provide opportunities for teachers to comment on their ratings. In particular, such comments allow teachers to offer their own perspective on student rating results and they can also provide context on any special circumstances surrounding a given course (e.g., new courses or innovations in teaching, a shift from an elective to a mandatory course, changes in departmental grading standards, and student resistance to certain types of material).

Faculty use SET summaries to diagnose their teaching performance and develop strategies for improvement. Frustrations arise when faculty are unable to use SET summaries to improve their teaching performance. Many teachers point out that the student written comments are more helpful when considering course changes, than quantitative feedback that only indicates the presence of student dissatisfaction. In order to make proper change to the course, teachers should both be motivated and able to make relevant adjustments. The ability to make relevant adjustments will usually increase as a result of participation in teacher training programs which will also encourage teachers to involve both students and peers in teaching development activities. Finally, it might be considered to encourage the teachers to use different kinds of consultations by faculty developers and/or peers to interpret the student feedback (ratings and comments) and discuss relevant measures to take.

6.2 Summary of Findings

This thesis, includes the five articles and a report presented in chapters 7- 12. It addresses different aspects of student course and teacher evaluation. The thesis consists of work combining various multivariate statistics tools and text-mining methods with applications to the education data. The results are presented through publications in both the field of applied statistics and the in the field of education.

In each problem, a set of data has been created and appropriate methods and tools have been applied. Finally, the methods have been applied to solve a problem, in each case providing new knowledge in the student evaluation field. The results of the work demonstrate a high potential of application of statistical and text-mining tools for analysis of student evaluations.

The main findings of the work are:

- A strong (around 0.7) association was found between how students evaluate the course and how students evaluate the teacher. Moreover, the relationship was found to be relatively stable over time for well established courses at DTU. Marsh (2007) indicated that students primarily evaluate the teacher rather than the course. However, according to findings of this thesis that neither of the first pair of canonical variables is a good overall predictor of the opposite set of variables. The proportions of variance explained are quite low for both evaluation of the course and evaluation of the teacher.
- The work shows that additional valuable information can be extracted from the students open-ended feedback. Such information can provide additional insight on why students are satisfied or dissatisfied with the courses for DTU administration in a process of analysing and comparing SET results of multiple courses. Moreover, the study also illustrates the problems, that arise when simple text-mining tools are applied to such short texts as students open-ended feedback are. Research on student open-ended feedback is relatively rare. Jordan (2011) as well as this thesis suggests that the student comments are related to the quantitative scores from course evaluations. Moreover, both works shows that text-mining methods can be used to extract meaningful information from students comments that can be used by university administration and department study boars during analysis of SET results.
- As the results of an additional mid-term evaluation it was found that the evaluations showed positive improvements over the semester for courses

where teacher had access to the results of mid-term evaluation, and negative improvements for those without access. This is in line with findings by Cohen (1980) who suggested that on average the mid-term evaluations had made a modest but significant contribution to the improvement of teaching. In particular, questions related to the student feeling that he/she learned a lot, a general satisfaction with the course, a good continuity of the teaching activities, and the teacher being good at communicating the subject was found to show statistically significant differences.

- Student evaluation non-response bias was found to be present in the mid-term and end-of-term evaluations. Female students, with high GPA, taking the course for the first time were more likely to participate in the course evaluation survey at both time points. Works by Avery et al. (2006); Kherfi (2011); Fidelman (2007) also found gender and GPA among predictors of non-response. However these studies investigated the non-response of just end of term evaluation. Additional analysis of SET scores showed that even though students with a high GPA had a higher probability to participate in the evaluation survey, the GPA itself had little effect on the SET scores. However, the grade obtained on the course was strongly positively related with both SET participation and SET scores. A lot of studies found that SET scores and obtained grades are positively correlated (Cashin, 1995).

The development of an automatic teaching quality assessment tool can improve the processes of analysing existing courses and provide ideas or fields for further course quality improvement.

6.3 Recommendations

Student evaluation of teaching is one of several mechanisms for assessing teaching quality at educational institutions. SET is an important part of an overall strategy for improving the courses, teaching and student learning, but it should not be the only data source for evaluating instruction.

Based on the investigations of this thesis some actions can be done to improve the current student evaluation analysis practices at the Technical University of Denmark:

- Under the current course evaluation and course registration system set up it is impossible to distinguish between students who dropped out of the

course at the beginning of the semester or at the end. However, Crews and Curtis (2011) in their suggestions for online course evaluation systems noted the importance of ensuring that when students withdraw from a course they are also dropped from the evaluation system. Addressing of this issue will provide more accurate course evaluation response rates. Additionally, a course drop-out rate can also be an indicator of course and teaching quality.

- Based on the investigations presented in chapter 11, it seems to be preferable to conduct midterm evaluations instead of the end-of term evaluations if one is concerned with an improvement of the courses over a semester. This is especially true for the project courses (courses where the learning objectives are more general, rather than specific). In a project courses it is easier for the teacher to make adjustments suitable for the particular group of students. Mid-term evaluation is able to capture both types of course issues: issues that can be addressed during the semester (microscale teaching activities, like quantity of examples) and also issues that can only be addressed at the next semester (macroscale/general course issues, like textbook quality or teaching assistants).

- On the other hand end-of-semester evaluations summarize students' overall satisfaction or dissatisfaction with the course, but at the point when it is too late to make adjustments in teaching in the current semester. Therefore, it might be beneficial to conduct a short end-of-term evaluation with very limited number of questions that focus on general course issues (like overall course quality, or whether the course objectives were reached). In order to obtain student feedback on the entire teaching and learning process, including the alignment of assessment of students' learning with course objectives and teaching activities, an end-of-term student evaluation should be performed after the final exams.

- In order to make comparisons of evaluation results between courses, the homogeneity of the students of the courses should be taken into account. Since student-specific and course-specific characteristics have an impact on how students evaluate the courses.

- Evaluations of the large courses with very diverse students cannot be directly compared with a medium sized course for students with similar background. In such cases, course evaluations of the same course by different groups of students (i.e. different study lines) can provide more adequate information. In order to keep student anonymity, the group size should be at least 5 students. Such kind of analysis can help to improve the course for students who take the course that is not on their core discipline.

- Students should be encouraged to answer the course evaluation questionnaires. This is especially important for the open-ended questions, that

historically have low response rates. Improvement in quality and quantity of student written feedback is crucial for the development of automated tool, that can extract important patterns from student comments and SET scores. Initiatives like e-mail reminders, advertisement of importance of should be taken to improve the evaluation response rates. Some of the teachers at DTU provide students with the time during the lecture/practical sessions to evaluate the course. Such initiatives usually result in higher response rates. Short end-of-course evaluation can be done right after the final exam.

6.4 Challenges for the Future

Ideally, part of the analysis of the students' evaluation of course and teaching quality can be automated. There are a lot of statistical tools that can be used to find out various relationships within the data. In addition text mining tools may provide a method of processing a large amount of unstructured text, such as open-ended student comments. This could provide institutions with relevant (and hopefully actionable) information that is useful to not just the teacher, but to the program and institution as well. This may not be feasible in the nearest future, but with improvement of the methods and the quality of the data, such a automated system could provide information to educational institutions that is currently inaccessible.

However, the educational institutions should also invest in the quality and quantity of the data, collected via evaluation surveys. First, there is the actual data collection, how, when, and where the data is collected. There are a number of ways to encourage students to complete the survey responses. This is especially important for open-ended questions, that historically have a very low completion rate. There is also evidence that surveys with fewer questions tend to have higher response rates.

Next, how will the data be used? There can be some concerns in many organizations among faculty, administration, and students. Therefore, educational institutions should be careful to develop a comprehensive policy regarding the use of evaluation data.

Part II

Contributions

Chapter 7

Canonical Correlation Analysis of course and teacher evaluations.

Authors: Tamara Sliusarenko[1] and Bjarne Kjær Ersbøll[1]

[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark.

Paper presented at the 2nd International Conference on Computer Supported Education, 7-10 April, 2010, Valencia, Spain

Abstract

At the Technical University of Denmark course evaluations are performed by the students on a questionnaire. On one form the students are asked specific questions regarding the course. On a second form they are asked specific questions about the teacher. This study investigates the extent to which information obtained from the course evaluation form overlaps with information obtained from the teacher evaluation form. Employing canonical correlation analysis it was found that course and teacher evaluations are correlated. However, the structure of the canonical correlation is subject to change with changes in teaching methods from one year to another.

7.1 Introduction

Teacher evaluations and overall course quality evaluations are widely used in higher education. Students usually submit their feedback about the teacher and the course anonymously at the end of the course. Results are usually employed to improve courses for future students and to improve the instructor's effectiveness. Many researchers have stated that student rating is the most valid and practical source of data on teaching and course effectiveness (McKeachie, 1997). Therefore, research on student evaluations is critical to make improvements in course construction and teaching methods.

Many studies have been done based on the data from student evaluation addressing the relationship between student rating and students achievement (Cohen, 1981; Abrami et al., 1997). The main conclusion is that the student's achievement is correlated with the student's evaluation of the teacher and the course (Cohen et al., 2003).

The purpose of this research is to investigate the degree of association between students' evaluation of the course and students' evaluation of the teacher. This is done using canonical correlation analysis, which is designed to investigate correlations amongst two sets of variables. The other question we are trying to address is whether this association is consistent over time.

7.2 Data and Methods

7.2.1 Data source and study sample.

This research is based on questionnaire data from course evaluations at the Technical University of Denmark (DTU). On-line course evaluation is performed a week before the final week of the course. This usually means the week 12 out of 13 weeks of teaching. Two samples of observations from the introductory statistics course taught by the same instructor in two subsequent years were analyses: 131 observations from autumn 2007 and 183 observations from autumn 2008.

The questionnaire at DTU consists of three parts: Form A contains questions about the course; Form B contains questions about teacher. Finally, form C contains three open questions; that gives the students the opportunity to write their feedback "What went well?"; "What did not go so well?"; "Suggestions for changes". This particular analysis us based on investigation of the relationship between Form A and Form B. Questions used in this research are presented in (Table 10.1) and (Table 7.2) respectively.

Table 7.1: Questions in Form A

ID	Question
A.1.1	I think I am learning a lot in this course
A.1.2	I think the teaching method encourages my active participation
A.1.3	I think the teaching material is good
A.1.4	I think that throughout the course, the teacher has clearly communicated to me where I stand academically
A.1.5	I think the teacher creates good continuity between the different teaching activities
A.1.6	5 points is equivalent to 9 hours per week. I think my performance during the course is
A.1.7	I think the course description's prerequisites are
A.1.8	In general, I think this is a good course

Each student has five possibilities to rate questions from 5 to 1, where 5 means that the student strongly agrees with the underlying statement and 1 means that the student strongly disagree with the statement. For question A.1.6 5 corresponds to "much less" and 1 to "much more", while for A.1.7 5 corresponds to "too low" and 1 to "too high".

Table 7.2: Questions in Form B

ID	Question
B.1.1	I think that the teaching gives me a good grasp of the academic content of the course
B.1.2	I think the teacher is good at communicating the subject
B.1.3	I think the teacher motivates us to actively follow the class
B.2.1	I think that I generally understand what I am to do in our practical assignments/ lab courses/ group computation/group work/project work
B.2.2	I think the teacher is good at helping me understand the academic content
B.2.3	I think the teacher gives me useful feedback on my work

7.2.2 Methodology

Canonical correlation analysis (CCA), introduced by Hotelling (1935, 1936), was performed to investigate the degree of association between tee evaluation of the teacher and the evaluation of the course. CCA is a convenient method to investigate what is common amongst two sets of variables in a linear sense, and than also be used to produce a model equation which relates two sets of variables. It has similarities with both multivariate regression the principal component analysis (Thompson, 1984).

The main idea behind CCA is to find canonical variables in the form of two linear combinations:

$$\begin{aligned} w_1 &= a_{11}x_1 + a_{21}x_2 + ... + a_{p1}x_p \\ v_1 &= b_{11}y_1 + b_{21}y_2 + ... + b_{q1}y_q \end{aligned} \tag{7.1}$$

such that the coefficient a_{1i} and b_{1i} maximize the correlation between two canonical variables w_i, aoe v_1. This maximal correlation between the two canonical variables ss called the first canonical correlation. The coefficients of the linear combinations are called canonical coefficients or canonical weights.

The method continues by finding a second set of canonical variables, uncorrelated with the first pair that has maximal correlation, which produces the second pair of canonical variables. The maximum number of canonical variables is equal to the number of variables in the smaller set. A likelihood ratio test was used to investigate statistical significance of canonical correlations.

7.3 Results

7.3.1 Evidence from the data

From the simple descriptive statistics presented in Table 7.3 it is evident that there is a difference in student rating between 2007 and 2008 in both parts: the course and the teacher evaluation.

Table 7.3: 2007 and 2008 sample descriptive statistics

	Autumn 2007		Autumn 2008	
Question	Mean	Standard Deviation	Mean	Standard Deviation
A.1.1	4.34	0.74	4.02	0.76
A.1.2	4.11	0.84	3.91	0.83
A.1.3	3.98	0.88	3.88	0.95
A.1.4	3.52	1.06	3.24	1.06
A.1.5	4.20	0.79	4.03	0.83
A.1.6	3.24	0.69	3.40	0.71
A.1.7	2.98	0.19	3.02	0.23
A.1.8	4.31	0.73	4.09	0.82
B.1.1	4.66	0.54	4.34	0.81
B.1.2	4.79	0.46	4.48	0.76
B.1.3	4.73	0.53	4.40	0.83

The highest rated course specific questions in both years about the course are A.1.1 “I think I am learning a lot in this course” and A.1.8 “In general, I think this is a good course.”, but the rating is lower in 2008 than in 2007. On average students rate both course and the teacher better in 2007 than in 2008. This difference may be explained by the fact that in autumn 2007 the course was taught in the way of normal lecturing, but in autumn 2008 it was also covered by video.

7.3.2 Autumn semester 2007

The first canonical correlation was found to be equal to 0.64. This gives an overall indication of the degree of association between teacher and course evaluation. It is the only canonical variable which is significant (p-value $< 0{,}0001$), which indicates that the two sets of variables are correlated in only one dimension.

The next question that arises is “how do we interpret the canonical variables?”.

Table 7.4: Canonical structure analysis of 2007 sample

Question	Standardized Canonical Coefficients	Canonical factor loadings	Canonical cross- loadings
A.1.1	0.22	**0.78**	**0.50**
A.1.2	0.01	**0.73**	0.47
A.1.3	-0.07	0.38	0.24
A.1.4	-0.02	0.37	0.24
A.1.5	**0.46**	**0.82**	**0.53**
A.1.6	-0.10	-0.04	-0.03
A.1.7	-0.08	-0.14	-0.09
A.1.8	**0.51**	**0.89**	**0.57**
B.1.1	**0.82**	**0.98**	**0.63**
B.1.2	0,12	0,78	0,50
B.1.3	0,14	0,71	0,45

To answer this question standardized canonical coefficients should be investigated. These coefficients are reported in the first column of Table 4. We can see that in the canonical variable of the course evaluation questions A.1.5 (I think the teacher creates good continuity between the different teaching activities) and A.1.8 (In general, I think this is a good course) have the highest weights. In the teacher related canonical variable question B.1.1 (I think that the teaching gives me a good grasp of the academic content of the course) is the most important.

Structure correlation coefficients, called canonical factor loadings, are also used to interpret the importance of each original variable in the canonical variables. Canonical factor loading is the correlation between the original variables and the canonical variables. Variables with high canonical factor loading should be interpreted as being a part of the canonical variable. The first set of loadings between course evaluation variables and their canonical variable are presented in the second column of Table 4. Questions A.1.5 and A.1.8 have the highest correlation with the course related canonical variable. However, questions A.1.1 (I think I am learning a lot on this course) and A.1.2 (I think the teaching method encourages my active participation) also have high canonical factor loadings. Question B.1.1 has the highest correlation with the teacher related canonical variable.

Next we look at the cross correlations between the original course evaluation variables and the canonical variables of the teacher evaluation variables presented in the third row of Table 4. We can see that questions A.1.5 and A.1.8 also have the highest cross-correlations with the teacher related canonical vari-

able, questions A.1.1 also has quite a high canonical cross-loading. Question B.1.1 has the highest cross-correlation with the course related canonical variable.

An overall conclusion that can be made is that the canonical correlation of 0.64 in the autumn 2007 introductory statistics course is mainly due to the relationship between the teachers ability to give a good grasp of the academic content of the course from one side and a good continuity between teaching activities in the course, good content of the course and good overall quality of the course on the other side.

7.3.3 Autumn semester 2008

As in the case of autumn semester 2007 only the first canonical correlation, equal to 0.71, appears to be significantly different from zero (p-value<0,0001).

Table 7.5: Canonical structure analysis of 2008 sample

Question	Standardized Canonical Coefficients	Canonical factor loadings	Canonical cross- loadings
A.1.1	**0.39**	**0.88**	**0.62**
A.1.2	**0.47**	**0.87**	**0.62**
A.1.3	-0.03	0.61	0.43
A.1.4	0.08	0.40	0.28
A.1.5	0.17	0.71	0.51
A.1.6	0.03	-0.09	-0.07
A.1.7	0.08	-0.04	-0.03
A.1.8	0.16	**0.76**	**0.54**
B.1.1	**0.43**	0.89	0.63
B.1.2	0.11	0.90	0.64
B.1.3	**0.55**	**0.94**	**0.67**

Analyzing the standardized canonical coefficients from the first column of Table 5 we can conclude that in the canonical variable of the course evaluation question A.1.1 (I think I am learning a lot on this course) and question A.1.2 (I think the teaching method encourages my active participation) are important. In the teacher related canonical variable questions B.1.1 (I think that the teaching gives me a good grasp of the academic content of the course) and B.1.3 (I think the teacher gives me useful feedback on my work) are important. Analysis of the canonical factor loadings, presented in the second and third columns of Table 5, shows that questions A.1.1, A.1.2 and A.1.8 have the highest cor-

relations with their canonical variable. We can also see that the same three questions have the highest cross-correlation with the teacher evaluation canonical variable. Question B.1.3 has the highest correlation and cross-correlation with the corresponding canonical variables.

An overall conclusion is that the canonical correlation of 0,71 in the autumn semester 2008 course is mainly due to the relationship between the teacher's ability to motivate the students and a good teaching method that encourages active participation in the course, good course content, and overall quality of the course. This difference can be explained by the change in teaching method from normal lectures in 2007 to combined lectures and video sequences, which could be replayed by the students, in 2008. This was reflected to a very high degree in the verbal comments in form C.

Examples of verbal comments from 2007 are very much focused on the teacher: "Good dissemination", "Teacher seems pleased with his course", "Engaged teacher", "Gives a really good overview", "Inspiring teacher". Examples of verbal comments from 2008 on the other hand to a very large extent are concerned with the new teaching method: "Good idea to record the lectures – useful for preparation for the exam", "The possibility of downloading the lectures is fantastic", "Really good course, the video recordings really worked well!"

7.4 Conclusions

This study analyses the association between how students evaluate the course and how students evaluate the teacher in two subsequent years, using canonical correlation analysis. This association was found to be quite strong in both years: higher in 2008 than in 2007. The structure of the canonical correlations appears to be different for these two years. This is accounted for by the change in teaching method used by the same teacher in the two different years: in 2007 it was normal lecturing, but in 2008 it was also covered by video - and the students really liked that. Therefore, question A.1.2 that concerns the teaching method has more impact on the correlation between course evaluation and teacher evaluation in 2008 than in 2007. In 2008 the teacher's motivation for the students to actively follow the class has major impact on the correlation between the teacher evaluation and the course evaluation instead of good academic grasp as in 2007.

Future Work

This paper is the early stage of comprehensive research on student evaluation at the Technical University of Denmark. Questions that would be addressed in future work include consistency of the evaluation in courses over time, across courses, and comparison of mandatory vs. elective courses.

The study will also investigate the relationship between students' achievements and students' rating of the teacher and the course (Ersbøll, 2010). Furthermore, investigation of whether student specific characteristics such as age, gender, years of education, etc have relationship with the student evaluation and achievement. Information from qualitative answers is also important, so some text-mining type methods will be used in order to utilize information from Form C.

Chapter 8

How do student evaluations of courses and of instructors relate?

Authors: Tamara Sliusarenko[1] , Line H. Clemmensen[1] and Bjarne Kjær Ersbøll[1]

[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark.

Preliminary results presented at the XXVth International Biometric Conference, 5-10 December, 2010, Florianopolis, Brazil

Paper submitted to Journal of Educational and Behavioral Statistics

Abstract

Course evaluations are widely used by educational institutions to assess the quality of teaching. At the course evaluations, students are usually asked to rate different aspects of the course and of the teaching. We propose to apply canonical correlation analysis (CCA) in order to investigate the degree of association between how students evaluate the course and how students evaluate the teacher. Additionally it is possible to reveal the structure of this association.

However, student evaluation data is characterized by high correlations between the variables (questions) and, sometimes, by an insufficient sample size due to low evaluation response rates. These two problems can lead to inaccurate estimates, non-generalizable and hardly interpretable results of CCA. Thus, this paper explores whether recently developed regularized versions of the CCA method can be used to address these weaknesses. Two versions of canonical correlation analysis that incorporate L_1 and L_2 norm regularizations, respectively, were applied. Both methods give results with increased interpretability over traditional CCA on the present student evaluation data. The method shows robustness when evaluations over several years are examined.

8.1 Introduction

Teacher evaluations and overall course quality evaluations are widely used in higher education. Students usually submit their feedback about the teacher and the course anonymously at the end of the course. The results are usually employed to improve the courses for future students and to improve the instructors' effectiveness.

The research on student evaluations is important to make improvements in course construction and teaching methods. Student evaluation of teaching (SET) is a very well documented and studied tool. It was first introduced in 1918 by Kilpatrick. Since then many issues in course evaluations have been discussed in the literature. An overview of research on student ratings of instruction by Marsh (1987) demonstrates that student ratings are multidimensional, quite reliable, reasonably valid, and a useful tool for students, faculty and university administrators.

Several studies on SET data investigate the relationship between student ratings and student achievements (Cohen, 1981; Feldman, 1989a; Abrami et al., 1997). The main conclusion is that a student's achievement is correlated with

a student's evaluation of the teacher and the course. Other often discussed issues are relationships between the SET scores and various student-specific, course-specific and instructor-specific characteristics (Marsh, 1987).

This paper analyses the student evaluations from another angle; by investigating the correlation between how students evaluate the course and how students evaluate the teacher. The objective is to analyze the degree of association between student evaluations of courses and student evaluations of teachers. As a subject we have chosen to study a single course over time.

8.2 Literature Review

The common method to investigate the correlation amongst two sets of variables is canonical correlation analysis (CCA), introduced by Hotelling (1935). CCA is a convenient method for investigating what is common amongst two sets of variables in a linear sense, and it can also be used to produce a model which relates the two sets of variables through linear combinations. The method has similarities with both multivariate regression and principal component analysis.

Application of CCA when variables in the sets are highly correlated or when the sample size is insufficient can lead to computational problems, inaccurate estimates of parameters or non-generalizable results. One way to deal with these problems is to introduce a regularization step into the calculations.

The first attempt to introduce the ridge regression technique, developed by Hoerl and Kennard (1970), to the problem of canonical correlation analysis was proposed by Vinod (1976) and later developed by Leurgans et al. (1993).

In the recent years, canonical correlation analysis has gained popularity as a method for the analysis of genomic data, which is characterized by the fact that the number of features generally greatly exceeds the number of observations. Therefore, some researchers have tried to develop a method that incorporates variable selection and produces linear combinations of small subsets of variables from each set of variables with maximal correlation.

The first development of Sparse CCA was presented in Parkhomenko et al. (2007), who proposed an iterative algorithm that uses soft thresholding for feature selection. This approach is related to sparse principal component analysis (Zou et al., 2006). Waaijenborg and Zwinderman (2007) adapted the elastic net (Zou and Hastie, 2005) to canonical correlation analysis. Various approaches to introduce sparsity into the CCA framework were proposed in more recent works

by Le Cao et al. (2009), Witten et al. (2009b), Hardoon and Shawe-Taylor (2011), Chen et al. (2012). Sparse CCA solves the problem of interpretability providing sparse sets of associated variables. These results are expected to be more robust and generalize better outside the particular study.

In cases when the relationship between the variables in the set is non-linear Kernel CCA proposed by Akaho (2006) can be used. Tenenhaus and Tenenhaus (2011) proposed the generalization of regularized canonical correlation analysis to three or more sets of variables. Golugula et al. (2011) presented a supervised modification to CCA and RCCA, which is able to incorporate a supervised feature selection scheme to perform regularization.

8.3 Methodology

8.3.1 Canonical Correlation Analysis

Canonical correlation analysis (CCA) was used to investigate the degree of association between the evaluation of the teacher and the evaluation of the course. CCA finds linear combinations of variables with the highest correlation between two sets of variables.

The method considers two matrices X and Y of order $n \times p$ and $n \times q$ respectively. The columns of X and Y correspond to variables and the rows correspond to experimental units. Classical CCA assumes $p \leq n$ and $q \leq n$. The main idea behind CCA is to find canonical variables in the form of two linear combinations:

$$\begin{aligned} w_1 &= a_{11}x_1 + a_{21}x_2 + ... + a_{p1}x_p \\ v_1 &= b_{11}y_1 + b_{21}y_2 + ... + b_{q1}y_q \end{aligned} \tag{8.1}$$

such that the coefficients a_{i1} and b_{i1} maximize the correlation between two canonical variables w_1, and v_1. In other words, the problem consists in solving

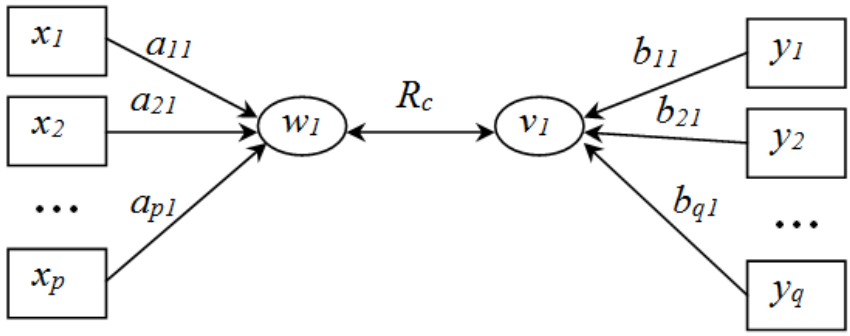
$$R_1 = corr\left(v_1, w_1\right) = \max_{a,b} corr\left(a^T X, b^T Y\right) \tag{8.2}$$

This maximal correlation between the two canonical variables v_1 and w_1 that are sometimes called canonical variates, is called the first canonical correlation. The coefficients of the linear combinations are called canonical coefficients or canonical weights.

The method continues by finding a second set of canonical variables, uncorrelated with the first pair that has maximal correlation. Wilks's lambda is used to test the significance of the canonical correlations.

Figure 8.1 illustrates the variable relationships in a hypothetical CCA. To answer the question "which variables are contributing to the relationship between the two sets?" the standardized canonical weights (i.e. coefficients used in linear equations that combine observed variables into latent canonical variable) and structure coefficients, also called canonical factor loadings, (i.e. correlations between observed variables and latent canonical variables) for the first significant canonical dimensions should be investigated (Thompson, 1984).

Figure 8.1: Visualization of CCA results



Canonical correlation analysis helps to identify the major association between student evaluations of the course and student evaluations of the teacher. To perform classical CCA the R package "CCA", developed Déjean and González (2009) was used. The package is freely available from the Comprehensive R Archive Network (CRAN) at `www.r-project.org`

8.3.2 Regularized Canonical Correlation Analysis

CCA cannot be performed when the variables $x_1, x_2, ..., x_p$ and/or $y_1, y_2, ..., y_q$are highly correlated. In this case the correlation matrices, that are used in the computational process, tend to be ill-conditioned and their inverses unreliable. To deal with this problem a regularization step can be included in the calculations. The principle of ridge regression (Hoerl and Kennard, 1970) was for the first time proposed to CCA by Vinod (1976) and then extended by Leurgans et al. (1993). It was also considered by González et al. (2009).

Ridge regression shrinks the weights by imposing a penalty on their size; highly correlated variables get similar weights, resulting in a grouping effect. For a linear regression ridge regression coefficients can be found using:

$$\hat{\beta}_{ridge} = arg\min_{\beta} |y - \mathbf{X}\beta|^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \tag{8.3}$$

In CCA the regularization is achieved by adding a corresponding identity matrix multiplied by a regularization parameter to the correlation matrices.

$$\begin{aligned} \Sigma_{XX}(\lambda_2^a) &= \Sigma_{XX} + \lambda_2^a I_p \\ \Sigma_{YY}(\lambda_2^b) &= \Sigma_{YY} + \lambda_2^b I_q \end{aligned} \tag{8.4}$$

As the result the matrices become nonsingular and a unique solution can be achieved. Regularized canonical correlation analysis (RCCA)is achieved by performing classical CCA where the correlation matrixes are substituted with $\Sigma_{XX}(\lambda_2^a)$ and $\Sigma_{XX}(\lambda_2^b)$. In order to choose "good" values of regularization parameters λ_2^a and λ_2^b , the k-fold cross-validation procedure can be used (González et al., 2009).

8.3.3 Sparse Canonical Correlation Analysis

Sparse CCA (SCCA) is an extension of CCA that can be performed when the number of observations is higher than the greatest amount of variables in both data sets ($n > max(p; q)$).In this case a selection of variables should be performed jointly with the analysis of the two data sets. SCCA can also help to solve the problem of interpretability providing sparse sets of associated variables. These results are expected to be more robust and generalize better outside the particular study.

Penalization methods such as lasso (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005) can be incorporated to canonical correlation analysis. Lasso is a penalization method developed Tibshirani (1996) that shrinks coefficients to zero. For a linear regression the Lasso regression coefficients can be found using:

$$\hat{\beta}_{lasso} = arg\min_{\beta} |y - \mathbf{X}\beta|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \tag{8.5}$$

Similar to regularized CCA, where ridge was introduced to solve the instability problem due to multicollinearity, it is possible to introduce the Lasso, which selects variables by setting the canonical correlation weights to zero.

Witten et al. (2009b) proposed the approach that introduces a new framework, named penalized matrix decomposition (PMA), for computing a rank K approximation for a matrix. The PMA results in a regularized version of the singular value decomposition (SVD), which is used to compute CCA results.

As was mentioned above, CCA finds the vectors a and b, that maximizes $corr(a^T X, b^T Y)$. The way to obtain penalized canonical variates is to impose L_1 penalties on vectors a and b. So the optimization problem can be written as:

$$\begin{aligned} &\max_{a,b} \left(a^T X^T Y b\right) \\ &s.t. \, \|a\|_2^2 \leq 1, \|b\|_2^2 \leq 1, \|a\|_1 \leq \lambda_1^a, \|b\|_1 \leq \lambda_1^b \end{aligned} \tag{8.6}$$

This problem can be solved using PMA approach. When λ_1^a and λ_1^b are small, some elements of a and b will be exactly zero. The algorithm yields sparse vectors a and b that maximize $cor(Xa, Yb)$. Values of regularization parameters λ_1^a and λ_1^b can be chosen using cross-validation.

To perform sparse CCA, the R package "PMA", written by Witten et al. (2009a) was used. The package allows to perform SCCA using the penalized matrix decomposition framework. It also contains a function that helps to select tuning parameters by using cross validation.

8.4 Data Description

Students at the University regularly evaluate courses by filling in web-forms a week before the final week of the course. The evaluations are intended to be a tool for quality assurance for: teachers, the department educational boards, and the department and university managements. On-line course evaluation at the university consists of three forms:

1. Form A contains specific quantitative questions about the course.
2. Form B contains specific quantitative questions about the teacher.
3. Form C gives students an opportunity to write their qualitative feedback.

This particular analysis is based on investigation of the relationship between answers in Form A and Form B. The questions can be seen in Table 8.1.

Table 8.1: The evaluation questions

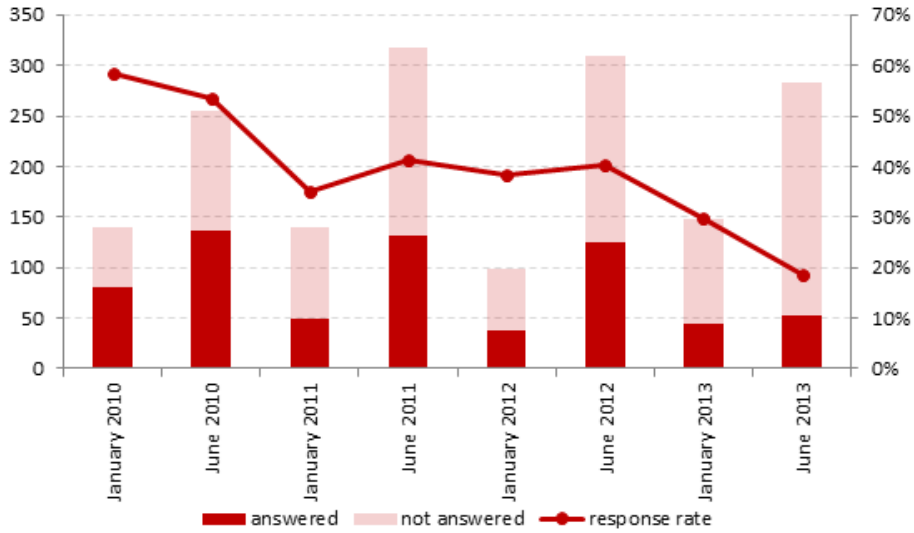
Id no.	Question	Short version (for reference)
A.1.1	I think I am learning a lot in this course	Learning a lot
A.1.2	I think the teaching method encourages my active participation	TM activates
A.1.3	I think the teaching material is good	Material
A.1.4	I think that throughout the course, the teacher has clearly communicated to me where I stand academically	Feedback
A.1.5	I think the teacher creates good continuity between the different teaching activities	TAs continuity
A.1.6	5 points is equivalent to 9 hours per week. I think my performance during the course is	Work load
A.1.7	I think the course description's prerequisites are	Prerequisites
A.1.8	In general, I think this is a good course	General
B.1.1	I think that the teacher gives me a good grasp of the academic content of the course	Good grasp
B.1.2	I think the teacher is good at communicating the subject	Communication
B.1.3	I think the teacher motivates us to actively follow the class	Motivate activity
B.2.1	I think that I generally understand what I am to do in our practical assignments/lab courses/group computation/group work/project work	Instructions
B.2.2	I think the teacher is good at helping me understand the academic content	Understanding
B.2.3	I think the teacher gives me useful feedback on my work	Feedback
B.3.1	I think the teacher's communication skills in English are good	English/English skills

The students rate the questions on a 5 point Likert scale (Likert, 1932) from 5 to 1, where 5 corresponds to the student "strongly agreeing" with the underlying statement and 1 corresponds to the student "strongly disagreeing" with the statement. For questions A.1.6 and A.1.7, a 5 corresponds to "too high" and 1 to "too low". In a sense for these two questions a 3 corresponds to satisfactory and anything else (higher or lower) corresponds to less satisfactory.

It is common practice to have just first 3 questions for the teacher evaluation (B.1.1-B.1.3) for large courses. In such cases, the second part of the form B (questions B.2.1-B.2.3) is active for the teaching assistants only. Here, we examine one course "Introductory Programming with Matlab". The course is one of the largest courses at the university where all 6 questions from the teacher evaluation (Form B) are usually active.

The Introductory Programming with Matlab course is available 4 times per year: twice as a 13-week course (fall and spring semesters) and twice as an intensive 3-week course (January and June). The numbers of students that follow the course are very different from semester to semester. Here we will focus on the intensive 3-week version of the course. June courses are more popular (approximately 300 students) than the January courses (around 100-150 students). Figure 12.1 shows the number of students registered for the course and the course evaluation response rate over the period from January 2010 to June 2013.

Figure 8.2: Number of course participants and evaluation response rate from January 2010 to June 2013.



For the comparison of methods we use results from one semester (January 2010), and for the robustness study we examine the same course at two other time points (June 2011 and June 2012). It should be noted, that students who participate in the course have very different backgrounds. Students at the university are obligated to take one programming course. Therefore the "Introductory Programming with Matlab" is a quite popular course among students on ‘non-

programming' study-lines.

8.5 Results

This section first presents a summary of the data, secondly it presents the results of three versions of the canonical correlation analysis performed on the same data. Finally, the results of the robustness study are presented.

8.5.1 Evidence from the data

The data set under investigation consists of 69 observations from the "Introductory Programming with Matlab" course held in January 2010. The course is one of the largest courses at the university, where the teacher is evaluated using all 6 questions from form B. Table 8.2 presents means and standard deviations of the answers from the responders. On average students gave rates below 3 to both the teacher and the course.

Table 8.2: Variable Mean and Standard Deviation

Evaluation of the course			Evaluation of the teacher		
Question	Mean	St. Dev.	Question	Mean	St. Dev.
A.1.1	2.46	1.16	B.1.1	2.57	1.10
A.1.2	2.11	1.09	B.1.2	2.80	1.30
A.1.3	2.62	1.18	B.1.3	3.01	1.29
A.1.4	2.43	0.98	B.2.1	2.22	1.08
A.1.5	2.58	1.03	B.2.2	2.50	1.11
A.1.6	2.67	0.83	B.2.3	2.42	1.05
A.1.7	3.06	0.45			
A.1.8	2.36	1.06			

Question A.1.2 (I think the teaching method encourages my active participation) got the lowest average grade among course-related questions and question B.2.1 (I think that I generally understand what I am to do in our practical assignments) got the lowest average grade among teacher-related questions.

Table 8.3 and Table 8.4 present the correlations between the variables from Form A and Form B.

The correlations appear to be quite high especially within Form B. This can lead to uninterpretable results of classical CCA.

Table 8.3: Correlations among the Form A variables

question	A.1.1	A.1.2	A.1.3	A.1.4	A.1.5	A.1.6	A.1.7	A.1.8
A.1.1	1.00							
A.1.2	**0.72**	1.00						
A.1.3	**0.57**	0.48	1.00					
A.1.4	0.34	0.24	0.44	1.00				
A.1.5	**0.55**	**0.54**	**0.53**	0.45	1.00			
A.1.6	-0.34	-0.21	-0.11	0.07	-0.16	1.00		
A.1.7	0.01	0.19	0.21	0.24	0.15	0.29	1.00	
A.1.8	**0.83**	**0.77**	**0.61**	0.37	**0.56**	-0.24	0.08	1.00

Table 8.4: Correlations among the Form B variables

question	B.1.1	B.1.2	B.1.3	B.2.1	B.2.2	B.2.3
B.1.1	1.00					
B.1.2	**0.81**	1.00				
B.1.3	**0.81**	**0.85**	1.00			
B.2.1	0.47	0.49	0.43	1.00		
B.2.2	**0.78**	**0.74**	**0.77**	**0.58**	1.00	
B.2.3	**0.64**	**0.57**	**0.67**	**0.55**	**0.78**	1.00

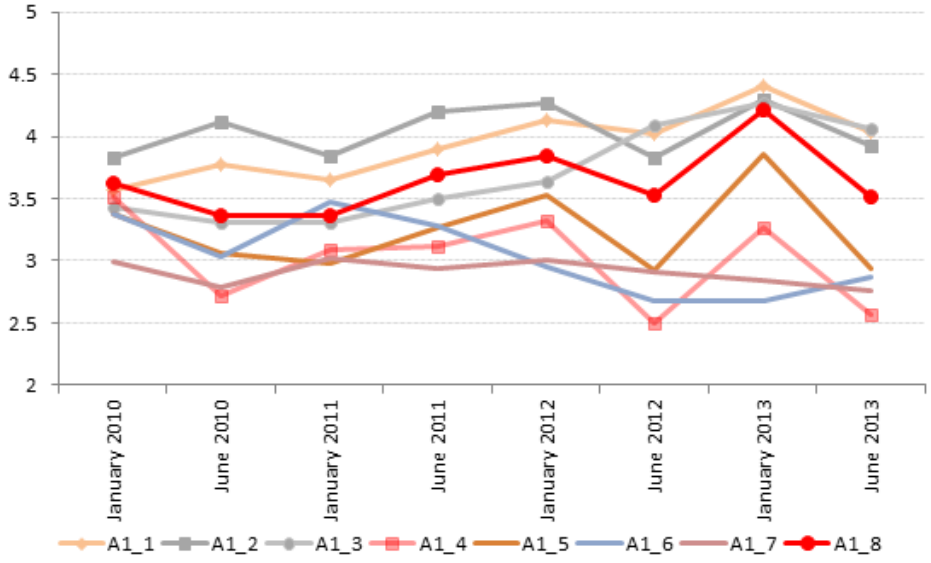
Figure 8.3 presents the average SET scores of the evaluation of the course (Form A) starting from January 2010 until June 2013.

There were some changes to the course during the period. In June 2010 the course was run by a new teacher, who introduced a new textbook, which seems much better than the Matlab notes used before as the course material got better feedback after this (A.1.3). Additionally, the course responsible team continuously works on improvement of the course and on making it less dependent on teacher and teaching assistants. Overall, there is a tendency of improvement of SET scores over the period from January 2010 to June 2013, with exception of the June 2012 semester, when the course got lower evaluation results.

8.5.2 CCA Results

Figure 8.4 presents the canonical correlations and corresponding p-values for the significance test of each canonical correlation. In general the number of canonical correlations is equal to the number of variables in the smallest set. However, the test shows that only the first 4 canonical correlations are statistically significant. This means that the structure of the association between course and teacher

Figure 8.3: Results of the evaluation of the course from January 2010 to June 2013



evaluations lies in 4 dimensions, which is hard to interpret. Values of canonical correlations give an overall indication of a strong association between teacher and course evaluation.

Table 8.5 presents the standardized canonical coefficients and table 8.6 presents structure canonical coefficients. These coefficients are used to find the structure of canonical correlation.

For the first canonical correlation, questions A.1.5 (continuity between the different teaching activities) and A.1.8 (overall course quality) from the course related questions are the most important. Among the teacher related, question B.1.1 (good grasp of the academic content) is the most important. However, due to high correlations between questions within each set of variables, canonical factor loadings indicate that the questions: A.1.1, A.1.2, A.1.3 from Form A and questions: B.1.2, B.1.3, B.2.2, B.2.3 from Form B are also important for the first canonical correlation. The structures of the other 3 canonical correlations can be found by similar analyses of corresponding coefficients.

The square root of the first canonical correlation shows the proportion of the

Figure 8.4: Canonical correlations and corresponding p-values

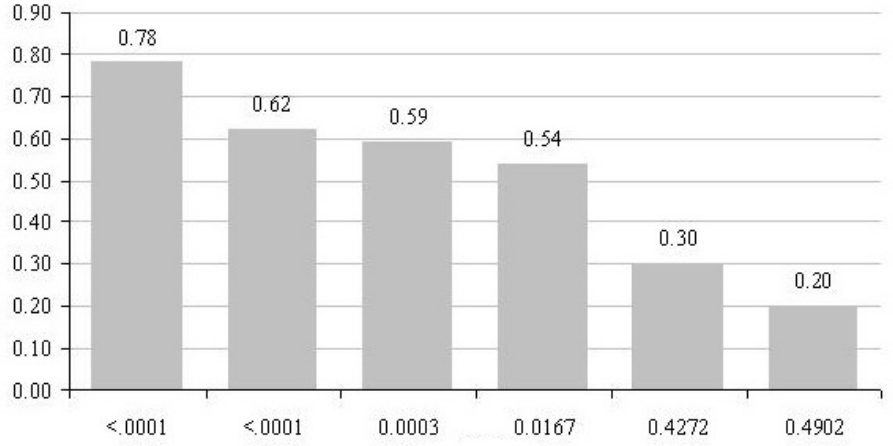


Table 8.5: Standardized canonical coefficients

Standardized Canonical Coefficients for the Form A variables					Standardized Canonical Coefficients for the Form B variables				
Form A	V1	V2	V3	V4	Form B	W1	W2	W3	W4
A.1.1	-0.03	-0.31	**0.86**	**-1.12**	B.1.1	**0.80**	**-0.86**	**0.65**	**-0.77**
A.1.2	-0.16	0.08	0.33	0.34	B.1.2	0.28	**1.37**	-0.33	**0.75**
A.1.3	0.34	**0.90**	-0.12	**-0.70**	B.1.3	0.18	-0.25	**-1.19**	-0.28
A.1.4	-0.10	**-0.50**	**0.78**	0.06	B.2.1	-0.03	**0.68**	**0.61**	-0.08
A.1.5	**0.60**	**-0.67**	**-0.54**	0.26	B.2.2	-0.17	-0.25	**0.59**	**-0.72**
A.1.6	-0.11	**0.09**	0.35	0.08	B.2.3	-0.09	-0.48	0.06	**1.55**
A.1.7	-0.12	**0.39**	0.16	0.39					
A.1.8	**0.42**	**0.38**	**-0.55**	**1.17**					

variance in the first canonical variate of one set of variables explained by the first canonical variate of the other set of variables. For the first canonical varite the proportion of explained variance is 61%.

The canonical redundancy analysis shows that neither of the first pair of canonical variables is a good overall predictor of the opposite set of variables, the proportions of variance explained being 0.24 and 0.35 for evaluation of the course and evaluation of the teacher respectively.

A four-dimensional structure of association between student evaluation of the course and student evaluation of the instructor can be a signal of data overfitting due to an insufficient sample size. Another problem is that correlations

Table 8.6: Structure canonical coefficients

Correlations between the Form A variables and their canonical variables					Correlations between the Form B variables and their canonical variables				
Form A	V1	V2	V3	V4	Form B	W1	W2	W3	W4
A.1.1	**0.73**	0.01	0.42	-0.16	B.1.1	**0.97**	-0.14	0.20	0.01
A.1.2	**0.60**	0.15	0.33	0.31	B.1.2	**0.89**	0.33	-0.06	0.20
A.1.3	**0.76**	**0.49**	0.24	-0.23	B.1.3	**0.87**	0.00	-0.21	0.21
A.1.4	0.38	-0.24	**0.71**	0.11	B.2.1	0.42	**0.43**	**0.62**	0.24
A.1.5	**0.87**	-0.28	0.05	0.19	B.2.2	**0.71**	-0.07	0.35	0.20
A.1.6	-0.35	0.17	0.32	0.25	B.2.3	**0.56**	-0.24	0.29	**0.69**
A.1.7	-0.02	0.43	0.37	**0.48**					
A.1.8	**0.79**	0.18	0.26	0.25					

between the variables within Form A and Form B are quite high. Therefore, the CCA results are hard to interpret. Dimension reduction methods such as regularized and sparse versions of canonical correlation analysis should be used to obtain easier interpretable results.

8.5.3 Regularized CCA Results

The regularization was achieved by adding to the correlation matrices a corresponding identity matrix multiplied by a regularization parameter as described in the methods section. The cross-validation procedure was used to find the optimal regularization parameters. Only the first canonical correlation, equal to 0.70, appeared to be statistically significant ($p - value = 0.025$). Thus, this canonical correlation structure has only one dimension, compared to the four-dimensional result of classical CCA. This results in a simpler and more generalizable model of the association between evaluation of the course and evaluation of the teacher.

The interpretation of the results of regularized canonical correlation analysis is similar to the interpretation of the results of classical CCA. To investigate the structure of the canonical correlation, the standardized canonical coefficients and the structure canonical coefficients (canonical factor loadings and canonical factor cross-ladings) reported in Table 8.7 should be analyzed.

The analysis of the standardized canonical weights shows that questions A.1.3 (teaching material is good), in addition to A.1.5 and A.1.8 seen in the classical CCA, from the course related questions are the most important variables.

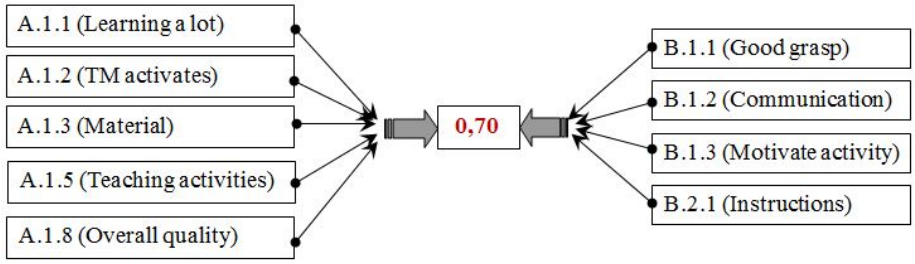
Table 8.7: Canonical weights and structure coefficients

	Evaluation of the course				Evaluation of the teacher		
	Standardized canonical weights	Canonical factor loadings	Canonical cross - loadings		Standardized canonical weights	Canonical factor loadings	Canonical cross - loadings
A.1.1	-0.08	**-0.81**	**-0.57**	B.1.1	**-0.53**	**-0.96**	**-0.75**
A.1.2	-0.03	**-0.72**	**-0.51**	B.1.2	**-0.33**	**-0.93**	**-0.69**
A.1.3	**-0.33**	**-0.82**	**-0.62**	B.1.3	-0.10	**-0.88**	**-0.63**
A.1.4	-0.03	-0.49	-0.34	B.2.1	-0,11	**-0.58**	-0.42
A.1.5	**-0.38**	**-0.83**	**-0.65**	B.2.2	-0.03	**-0.82**	**-0.58**
A.1.6	0.05	0.28	0.21	B.2.3	-0.03	**-0.66**	-0.46
A.1.7	0.03	-0.12	0.05				
A.1.8	**-0.29**	**-0.85**	**0.65**				

Among the teacher related questions, B.1.1 (teacher gives me a good grasp of the academic content) and B.1.2 (teacher is good at communicating the subject) are the most important. An analysis of the canonical factor loadings and the cross-loadings shows that A.1.1 and A.1.2 from Form A and questions B.1.3 and B.2.2 from Form B also contribute to the canonical correlation.

Figure 8.5 presents the variables from Form A and Form B that contribute to the latent canonical variables.

Figure 8.5: RCCA: Questions that contribute to canonical correlation



An overall conclusion that can be made is that the correlation of 0.70 in the "Introductory Programming with Matlab" course is mainly due to the relationship between the content of the course, the teaching methods, the continuity between teaching activities in the course, the teaching material and the overall quality of the course from one side and the teachers ability to give a good grasp

of the academic content of the course, the teachers ability to motivate the students, the teachers communication about the subject and the understanding of practical assignments on the other side.

8.5.4 Sparse CCA Results

The first canonical correlation of the sparse CCA was found to be equal to 0.75, which is the correlation between a linear combination of 4 variables from Form A and a linear combination of 3 variables from Form B. Table 8.8 presents the coefficients that correspond to these linear combinations.

Table 8.8: Sparse canonical coefficients

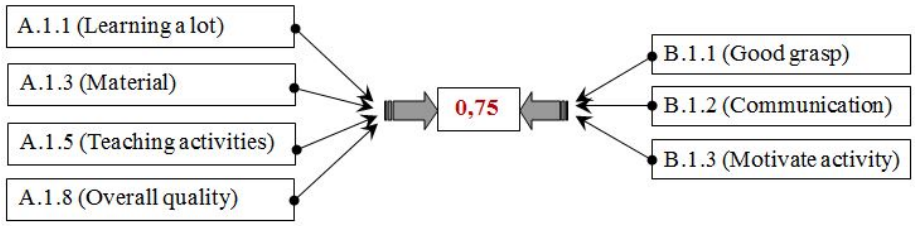
Evaluation of the course		Evaluation of the teacher	
question	coef.	question	coef.
A.1.1	-0.08	B.1.1	-0.94
A.1.2	0	B.1.2	-0.32
A.1.3	-0.17	B.1.3	-0.14
A.1.4	0	B.2.1	0
A.1.5	-0.83	B.2.2	0
A.1.6	0	B.2.3	0
A.1.7	0		
A.1.8	-0.53		

From Form A, the questions: A.1.1, A.1.3, A.1.5 and A.1.8 contribute to the course related latent canonical variable while from Form B, the questions: B.1.1, B.1.2 and B.1.3 contribute to the teacher related latent canonical variable. This model is also simpler than the one obtained from classical CCA. Furthermore, it also involves less variables than the model obtained from the regularized version of CCA (it does not contain questions A.1.2, A.1.4, A.1.6, A.1.7, B.2.1, B.2.2, and B.2.3).

Figure 8.6 presents the variables from Form A and Form B that contribute to the latent canonical variables.

The conclusion is that the canonical correlation of 0.75 in the "Introductory Programming with Matlab" course is mainly due to the relationship between the good continuity between teaching activities in the course, content of the course, teaching material and overall quality of the course from one side and teachers ability to give a good grasp of the academic content of the course, teachers ability to motivate the students and teachers good communication about the subject on the other side.

Figure 8.6: SCCA: Questions that contribute to canonical correlation



8.5.5 Stability of the results

To check for the stability of the correlation structures, subsequent years of the course should be analyzed. Figure 8.7 and figure 8.8 present the canonical correlation structures for the association between evaluation of the course and evaluation of the teacher in June 2011 and June 2012, respectively.

Figure 8.7: The structure of canonical correlation between the two parts of course evaluation in June 2011

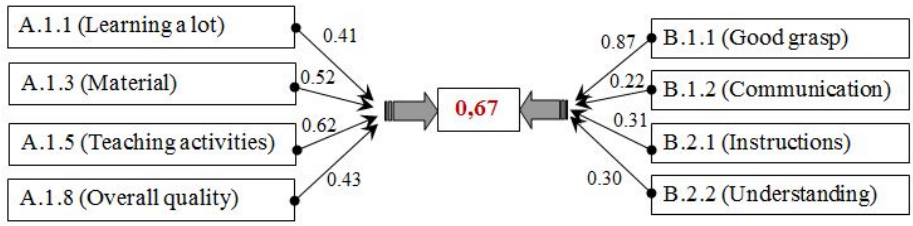
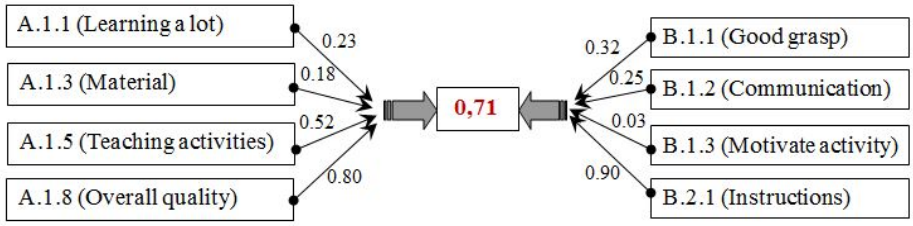


Figure 8.8: The structure of canonical correlation between the two parts of course evaluation in June 2012



Overall, the two structures are similar. The only difference was on the evaluation

of the teacher, where question B.2.2 (The teacher is good at helping me to understand the academic content) form the structure in 2011, while B.1.3 (The teacher motivates us to actively follow the class) was in the canonical correlation structure for 2012. Figures also show the weights, each variable had in the latent canonical variable. The weights were different in the two years. However the changes in the canonical correlation structures can be explained by the fact that the main teachers of the course for all three semesters (January 2010, June 2011 and June 2012) were different.

8.6 Discussion

The study have found that association between how students evaluate the course and how they evaluate the teacher of the course is strong (correlation is around 70 %), and the structure of this association is relatively stable over time. The square root of the first canonical correlation shows the proportion of the variance in the first canonical variate of one set of variables explained by the first canonical variate of the other set of variables (around 50%).

Having this strong relationship, better courses and therefore better SET results can be achieved in several different ways: improvement in a course can lead to better evaluation of teacher, and improvement of the teacher qualities, can lead to better evaluation of the course. However, Marsh (2007) indicated that students primarily evaluate the teacher rather than the course. But there is still some 30% left of this dimension, and there are the orthogonal dimensions as well.

The canonical redundancy analysis for the traditional CCA shows that neither of the first pair of canonical variables is a good overall predictor of the opposite set of variables, the proportions of variance explained being 0.24 and 0.35 for evaluation of the course and evaluation of the teacher respectively. There is no guarantee that what students answer on the course evaluation is not a function of the teacher, but the students do have some parts of non-correlating responses for the two parts of evaluation although there is also a strong association.

In case the structure of association between evaluation of the course and evaluation of the teacher is stable, SET administrators might consider to reduce the number of questions in the questionnaire in order to gain better response rates. However, that should be done very carefully. SETs must be multidimensional, in order to reflect multidimensionality of such a complex activity as teaching. According to Marsh and Roche (1997) , the strongest support for the multidimensionality of SETs is based on the nine factors: Learning/Value,

Instructor Enthusiasm, Organization/Clarity, Group Interaction, Individual Report, Breadth of Coverage, Examinations/Grading, Assignments/Readings, and Workload/Difficulty. The questionnaire, currently used at DTU is already small, but an analysis similar to this could be used by other educational institutions.

8.7 Conclusions

This study analyzed the association between how students evaluate a course and how students evaluate a teacher using canonical correlation analysis (CCA). Data from student evaluations is characterized by high correlations between the variables within each set of variables, therefore two modifications of the CCA method; regularized CCA and sparse CCA, together with classical CCA were applied to find the most interpretable model of association between the two evaluations.

The association between how students evaluate the course and how students evaluate the teacher was found to be quite strong in all three cases. However, applications of regularized and sparse CCA to the present student evaluation data give results with increased interpretability over traditional CCA.

The simplest model was obtained from sparse canonical correlation analysis, where an association between how students evaluate the course and how students evaluate the teacher was found to be due to the relationship between the good continuity between teaching activities in the course, the content of the course, the teaching material, and the overall quality of the course from the course side; and teachers ability to give a good grasp of the academic content of the course, the teachers ability to motivate the students and the teachers good communication about the subject on the teacher side.

Analysis of subsequent evaluations of the same course showed that the association between how students rate the teacher and the course was found to be subject to subtle changes with the change of teaching methods and with the change of instructor. These changes in the correlation structure were seen on the instructor side and not on the course side.

Chapter 9

Clustering the students comments

Author: Tamara Sliusarenko[1]

[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark.

Report for Summer school Matrix Methods for Data Mining and Pattern Recognition

August 23 - 27, 2010, DTU

9.1 Objective

Teacher evaluations and overall course quality evaluations are widely used in higher education. Results are usually employed to improve courses for future students and to improve the instructor's effectiveness. At the Technical University of Denmark (DTU) Students regularly evaluate the courses by filling in the so-called: "final-evaluation" web-forms on the intranet CampusNet.

The on-line based course evaluation is performed one week before the final week of the course and consists of three forms: Form A and Form B contains 5 point Likert scale (Likert, 1932) questions about the course and about the teacher respectively, while Form C gives the possibility of more qualitative answers on 3 questions: What went well?; What did not go so well?; Suggestions for changes;

In addition to analysis of quantitative answers for questions in Form A and Form B, it is interesting to analyze student's written comments in order to to improve analysis of student's evaluation at DTU. Teachers can react better and in a more productive way on a specific points of satisfaction or dissatisfaction from written comments then on the average quantitative rating of their work. Therefore, an application of some text mining tools to analysis of written comments, that can help to find some common patterns in student's comments, will help to improve courses quality and teachers effectiveness.

Comparing the patterns in comments for two or more subsequent courses performed by the same teacher we can answer the question whether instructor react on comments or not.

9.2 Methods

Text mining involves extracting information from unstructured data. Most of the methods are based on construction of word-document matrix which is highly-dimensional and sparse.

9.2.1 Latent Semantic Indexing

Latent Semantic Indexing (LSI) is a tool, that discovers semantically related words and phrases (Deerwester, 1988). The method is based on the principle that words that are used in the same contexts tend to have similar meanings.

Search engines, Internet Marketing Professionals, and Website Designers often use LSI in their day-to-day activities.

LSI uses singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts.

LSI begins by constructing a term-document matrix, to identify the occurrences of the unique terms within a collection of documents. Term-document matrix is usually very large and very sparse. Then a rank-reduced SVD is performed on the matrix to determine patterns in the relationships between the terms and concepts contained in the text. Efficient LSI algorithms only compute the first k singular values and term and document vectors as opposed to computing a full SVD and then truncating it (Langville, 2005). Retrieval is then performed using the database of singular values and vectors obtained from the truncated SVD.

9.2.2 Term-document matrix

A lot of the text mining methods are based on construction of a term-document matrix, high-dimensional and sparse mathematical matrix that describes the frequencies of terms that occur in a collection of documents. There are various ways to determine the value that each entry in the matrix should take, one of the most popular are term frequency and term frequency - inverse document frequency (tf-idf).

Tf-idf is a numerical value which reflects importance of a word for a document in a collection of documents.It increases proportionally to the number of times a word appears in the document, but with an offset by the frequency of the word in the collection of analyzed documents.

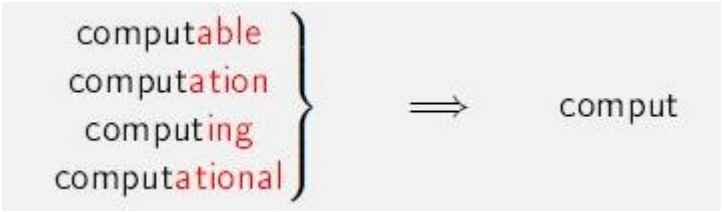
9.2.3 Text preprocessing

Texts preprocessing is an important part of natural language processing. The process usually starts with removing away useless characters like HTML tags, (sometimes) numbers, and punctuations, and continues with more optional stemming and stop word removal.

9.2.3.1 Stemming

Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form (Figure 9.1). The stem is not necessarily identical to the morphological root of the word.

Figure 9.1: Stemming



Stemmer developed by Porter (1980) is the most widely used and became *defacto* the standard algorithm used for English stemming.

9.2.3.2 Stop-word removal

The idea of stopword removal is to remove words that occur in "all documents'. The most frequent words often do not carry much meaning. For example words "the", "a", "of", "for", "in", etc. The concept was first introduced b in 1958 by Luhn (1958).

The beginning of stopword list built by Gerard Salton and Chris Buckley from Cornell University: *a, a's, able, about, above, according, accordingly, across, actually, after, afterwards, again, against, ain't, all, allow, allows, almost, alone, along, already, also, although, always, am, among, amongst, an, and, another, any, anybody...*

This wordlist that consist of 571 words in length, is freely available (Buckley and Salton).

9.2.4 Clustering

One of the drawbacks of SVD is that it is computationally difficult to update for new terms and new documents. In text mining clustering is often used as

alternative to SVD to group together records with similar words or words with similar meanings.

Many different techniques for clustering exist. One of the most common methods is k-means clustering, developed by MacQueen (1967). In general, the k-means method will produce exactly k different clusters of greatest possible distinction.

The best number of clusters k leading to the greatest separation (distance) is not known as *apriori* and must be computed from the data. A statistical measure of dissimilarity between records is used to separate records that are the most dissimilar and group together records that are the most similar. Usually, as the result of a k-means clustering analysis, means for each cluster on each dimension, usually called centroids, are analyzed to assess how distinct the clusters are.

9.3 Data Description

Two samples of observations from the Introductory Statistics course at DTU taught by the same instructor in two subsequent years were analyzed: In autumn 2007 131 students fill out the evaluation form, while in autumn 2008 183 students have responded.

The response rate on the open-ended questions is usually lower than on quantitative questions. Students has an opportunity to write their qualitative feedback in Form C, that consist of 3 questions:

C.1.1 What went well?

C.1.2 What did not go so well?

C.1.3 Suggestions for changes?

The Introductory Statistics course appeared to be good in both years, therefore there where not many comments for questions C.1.2 and C.1.3. Therefore comments for question C.1.1 (What went well?) are analyzed. The samples consist of:

- 71 comments for fall 2007
- 90 comments for fall 2008

The length of the comments ranges from 1 word, which is usually word "nothing", to a paragraph with 90 words.

Analysis of a good course in this case is still interesting because there was a change in teaching method introduced in 2008. The instructor changed his teaching method from normal lectures in 2007 to combined lectures and video sequences, which could be replayed by the students, in 2008.

Examples of verbal comments from 2007 are very much focused on the teacher: *"Good dissemination", "Teacher seems pleased with his course", "Engaged teacher", "Gives a really good overview", "Inspiring teacher".*

Examples of verbal comments from 2008 on the other hand to a very large extent are concerned with the new teaching method: *"Good idea to record the lectures – useful for preparation for the exam", "The possibility of downloading the lectures is fantastic", "Really good course, the video recordings really worked well!"*

9.4 Results

9.4.1 Stemming and removing of useless words

One of the problems of analysis of student's comments in DTU is that some comments are in Danish and some of them are in English. In order to deal with the problem all the comments, written in Danish, were translated into English via on-line Google translate, free statistical multilingual machine-translation service provided by Google Inc. Other problems of student comments are slang language, abbreviations and all types of typo mistakes.

Example of comment from fall 2007 sample:

Comment in Danish: *"Det er meget fint med løsningerne som bliver lagt ud. Det hjælper formidabelt meget ved løsningen af opgaverne derhjemme. For mig giver det en meget god indsigt og forståelse af løsningsmetoderne. Dog kunne notationen tit følge den i litteraturen lidt bedre, da det for en usagkyndig kan være svært at forstå, hvis ikke det ligner det man kan se i litteraturen."*

Comment translated into English: *"It is very fine with solutions which will be posted It helps tremendously much in solving the tasks at home For me it makes a very good insight and understanding of solution methods However*

could often follow the notation in the literature a little better since for an improper can be difficult to understand if not it looks like it can be seen in the literature"

After translation all the comments have been pre-processed: punctuation has been removed, all the words have been lowercased and stemmed, irrelevant and useless words has been removed.

Comment after stemming: *"it is veri fine with solut which will be post it help tremend much in solv the task at home for me it make a veri good insight and understand of solut method howev could often follow the notat in the literatur a littl better sinc for an improp can be difficult to understand if not it look like it can be seen in the literatur"*

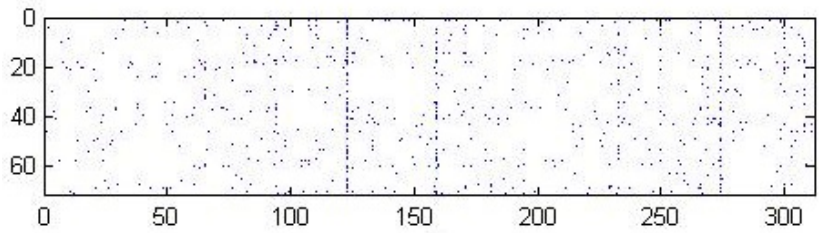
Comment after removing of useless words: *"fine with solut post help tremend solv task home veri good insight understand solut method often follow notat literatur littl better sinc improp difficult understand look seen literatur"*

9.4.2 Term-document matrices

Term-document matrix is usually very sparse.

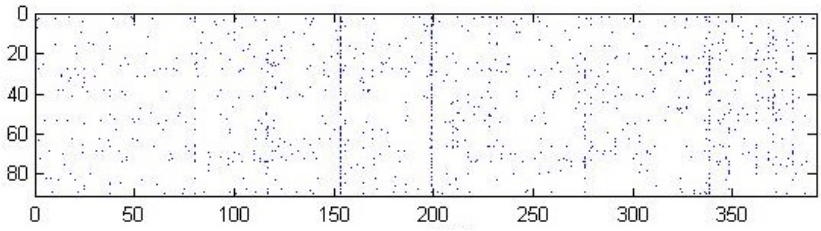
For 2007 figure 9.2 plots the 71x311 term-document matrix of students positive comments. The 104 words from stop word list have been removed. The sparsity coefficient is 3,07%.

Figure 9.2: Sparse term - document matrix for 2007



For 2008 Figure 9.3 plots the 90x391 term-document matrix of students positive comments. The 128 words from stop word list have been removed. The sparsity coefficient is 2,92%.

Figure 9.3: Sparse term - document matrix for 2008



9.4.3 K-means clustering results

To generate the therm-document matrices the MATLAB Toolbox Text to Matrix Generator (TMG) (Zeimpekis and Gallopoulos, 2005) has been used.

K-means separates documents into predetermined number of groups. The dominating words of means (centroids) for 4 clusters are presented below in Table 9.1.

Table 9.1: Dominant words in clusters for collection of comments for 2007

Centroid 1	Centroid 2	Centroid 3	Centroid 4
16 comments	31 comments	10 comments	14 comments
22,5%	43,7%	14,1%	19,7%
'good'	'course'	'convey'	'example'
'lecture'	'good'	'good'	'good'
'really'	'lecture'	'subject'	'help'
'teacher'	'well'	'super'	'lecture'
		'teacher'	'slide'
			'teacher'

In 2007 k-means algorithm performed quite well and separated documents by 4 main topics.

Cluster 1 refers to overall good quality of both the teacher and the course

Cluster 2 refers to good quality and content of the lectures and the course in general. This is the largest cluster.

Cluster 3 refers to teachers ability to convey the subject

Cluster 4 refers to good teaching assistants. Since Danish word "hjælplerer" is translated as "help teacher" by Google translate. Moreover the names

of two teacher assistants are among frequent words in this cluster.

The result of clusterization of 90 comments from fall 2008 is a little bit different. Dominating words of means for 4 clusters are presented in Table 9.2.

Table 9.2: Dominant words in clusters for collection of comments for 2008

Centroid 1	Centroid 2	Centroid 3	Centroid 4
14 comments	21 comments	28 comments	27 comments
15,6%	23,3%	31,1%	30,0%
'really'	'lecture'	'well'	'lecture'
'lecture'	'teacher'	'good'	'teacher'
'teacher'	'really'	'course'	'example'
'note'	'well'	'video'	'help'
'task'		'idea'	
'well'			
'course'			
'solutions'			

Dominating words in clusters for 2008 are similar to those in 2007.

Cluster 1 refers to overall good quality of the teacher, course, lecture notes and tasks and solutions

Cluster 2 refers to good quality of the teacher and the course. Almost the same dominant works as in cluster 1 for 2007 sample.

Cluster 3 refers to the change in the teaching method with video recorded lectures posted on the course web-page.

Cluster 4 refers to good the good teaching assistants. Dominant words are almost the same as in cluster 4 for 2007.

Comparing the results of k-means clusterization for two subsequent years of the same course, it can be concluded that changes in teaching methods are reflected in students open-ended comments from course evaluations. The whole new cluster the reflect the introduction if video-recorded lectures was identified in fall 2008 Introductory Statistics course. The smallest cluster 3, from year 2007 comments, about teachers ability to convey the subject did not appear as the separate cluster in year 2008 comments.

9.4.4 SVD on term-document matrix

Under Latent Semantic Indexing (LSI) methodology, the Singular Value Decomposition (SVD) method can be used to cluster documents and carry out information retrieval by using concepts as opposed to exact word-matching. Concepts are usually reflected in the words or phrases. However,

- One term may have multiple meaning
- Different terms may have the same meaning (for example, words 'teacher' and 'lecturer' are almost the same in context of students feedback on teacher and course quality)

LSI applies the truncated version of SVD to a term-document matrix. It first decomposes term-document matrix by SVD, and then approximating the matrix using first k finding singular vectors. The idea of truncation helps to

- Capture the most important relationships in the term-document matrix
- Ignore unimportant relationships
- Rebuilt the matrix using only important features

The method is useful only if number of the first singular vectors used in truncated SVD k is much less the the rank of the term-document matrix. If k is too large, it doesn't capture the underlying latent semantic space; while if k is too small, too much information is lost. The is no principled way of determining the best k.

Term-Document matrices are usually high dimensional and sparse (figure 9.4 and figure 9.5). The Singular Value Decomposition (SVD) destroys sparsity by finding singular vectors that are orthogonal to each other.

Figure 9.4 shows plots the singular values for SVD for term – document matrix of positive comments for 2007. One of the challenges to LSI is determining the optimal number of dimensions to use for performing the SVD. Research has demonstrated that around 300 dimensions will usually provide the best results with moderate-sized document collections (hundreds of thousands of documents) and perhaps 400 dimensions for larger document collections (millions of documents) (Bradford, 2008), however collection of students open-ended feedback is much smaller.

Figure 9.4: Singular values for SVD for term – document matrix for 2007

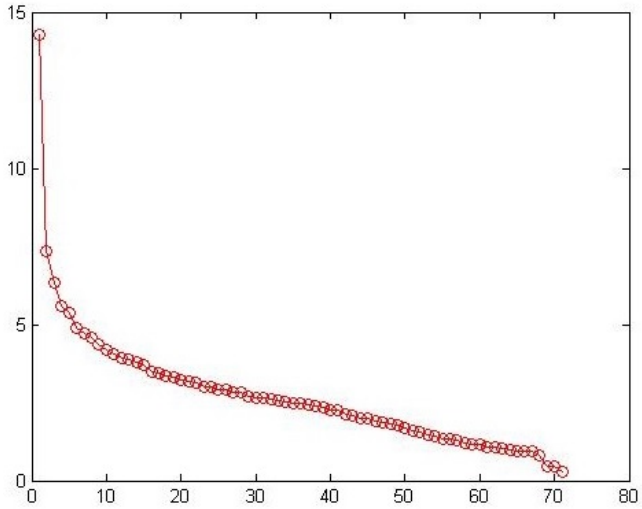


Table 9.3 shows the dominant words (concepts) in first 8 singular vectors. Singular vectors are orthogonal to each other, however some of the words appear to be dominant in two or more vectors. This is mainly due to different meaning of the same concept, when used in different contexts.

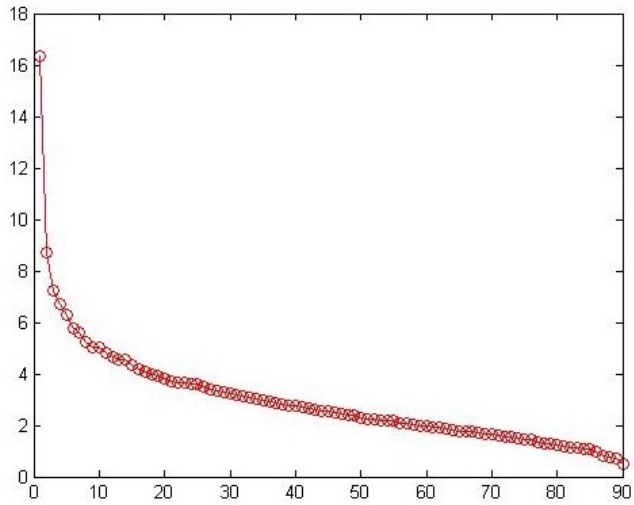
Table 9.3: dominant words in first 8 singular vectors for 2007 sample

v. 1	v. 2	v. 3	v. 4	v. 5	v. 6	v. 7	v. 8
example	good	course	course	assignment	Donald	course	example
good	lecture	good	exam	course	example	example	excel
lecture	super	lecture	read	example	lecture	good	fine
really	teacher		really	slides	really	many	generate
teacher			teacher	super	subject	read	overview
			well	well	well	really	read
						subject	statistics
						well	subject
							super
							understand

The singular vectors obtained cam be used for further query matching, when the new comments arrive. However, due to changes, the teacher makes from year to year, the the basis of singular vectors obtained in one semester may not be relevant for subsequent years.

Figure 9.5 shows plots the singular values for SVD for term – document matrix of positive comments for 2008 and table 9.4 shows the dominant words in first 8 singular vectors.

Figure 9.5: Singular values for SVD for term – document matrix for 2008



There is a difference between singular vectors for 2007 and 2008 samples. This is mostly due to introduction of video recording of the lectures; the concept 'video' is among dominant words in all singular vectors except vectors 4 and 7. Many students pointed out that video lectures, that were available on the Internet, helped them a lot during preparation of exams and home assignments in addition to compliments to the teacher or course content.

9.5 Conclusions

The report presents results of short tryout study of students qualitative comments for course evaluation at Technical University of Denmark. The study tries to apply two clustering methods, k-means and SVD, to cluster the open-ended feedback on a well established Introductory statistics course in two subsequent years. It appeared that available data came from one of the best courses in DTU that has not many drawbacks. Therefore question of weather teachers react on students written comments and suggestions is left unanswered.

The conclusion that can be made from given work is that students react on changes in teaching methods. In this study it is shown that introduction of

Table 9.4: dominant words in first 8 singular vectors for 2008 sample

v. 1	v. 2	v. 3	v. 4	v. 5	v. 6	v. 7	v. 8
good	good	lecture	course	course	example	course	course
lecture	lecture	really	good	example	good	exam	group
really	really	teacher	group	good	group	extra	help
teacher	teacher	understand	task	help	help	lesson	lot
video	video	video	teacher	lecture	lot	many	solutions
well		well	theory	lot	really	really	teach
			understand	really	teacher	record	think
			well	super	theory	room	under-stand
				teacher	think	slide	video
				think	understand	task	
				transmission	video	teacher	
				unde-stand		think	
				video		video	
				well		well	

video lectures is reflected in students comments. So it can be expected that analysis of whether instructor improves his/her course through the years can be done by analysis of written comments. Moreover such a significant changes cause changes in the basis vectors.

9.6 Future Work

The study has some questions left unanswered. For example it would be interesting to apply similar methods to analysis students of written comments of course that has some problems.

Other improvements that clearly can be done compared to with this work:

- Synonyms, such as "good" or "nice", or words that are used in particular phrases where they denote unique meaning can be combined for indexing
- Stemming and synonyms are highly language dependent operations. Therefore, support for different languages, in particular case Danish language, is extremely important.
- Introduction of some weighting or hierarchy organization of comments can help to improve results.

Chapter 10

Text mining in students' course evaluations: Relationships between open-ended comments and quantitative scores

---

Authors: Tamara Sliusarenko[1], Line H. Clemmensen[1] and Bjarne Kjær Ersbøll[1]

[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark.

Paper presented at the 5th International Conference on Computer Supported Education, 5-8 May, 2013, Aachen, Germany

10.1 Abstract

Extensive research has been done on student evaluations of teachers and courses based on quantitative data from evaluation questionnaires, but little research has examined students' written responses to open-ended questions and their relationships with quantitative scores. This paper analyzes such kind of relationship of a well established course at the Technical University of Denmark using statistical methods. Keyphrase extraction tool was used to find the main topics of students' comments, based on which the qualitative feedback was transformed into quantitative data for further statistical analysis. Application of factor analysis helped to reveal the important issues and the structure of the data hidden in the students' written comments, while regression analysis showed that some of the revealed factors have a significant impact on how students rate a course.

10.2 Introduction

Teacher evaluations and overall course quality evaluations, where students submit their feedback about the teacher and the course anonymously at the end of the course or during the course, are widely used in higher education. The results of such evaluations is one of the most common tools used by universities to improve courses for future students and to improve teachers effectiveness (Seldin, 1999; Wright, 2006). At the same time, student ratings is also one of the most controversial and highly-debated measures of course quality. Many have argued that there is no better option that provides the same sort of quantifiable and comparable data on teaching and course effectiveness (Abrami, 2001; McKeachie, 1997).

In addition to analysis of quantitative answers for questions, there is a need for analyzing students' written comments. Many instructors say that they get much more relevant information from students' written comments than they do from the quantitative scores. Teachers can use insights from the ' written feedback to make adjustments to future classes in a more productive way.

Student's written feedback is also of interest for university administration and study board, however it is hard to go trough all the comments from all courses taught at the university every semester. For the university administration and study board it is more convenient to have a general overview of the main points of student satisfaction and dissatisfaction, extracted from students written feedback.

A tool, that helps to automatically extract important points from open-ended questions from course evaluation, can add important information to the process of analysis and improvement of courses. This study is just an early stage that tries to find the most important patterns in students' written positive and negative feedback for one well established course, at the Technical University of Denmark (DTU) using simple statistical and text-mining tools.

10.3 Literature

Analysis of open-ended students' comments is problematic, because written comments have no built-in structure. Another challenge is that open-ended questions have much lower response rates than quantitative questions and there are some comments like "no comments" or "nothing", that are unhelpful. On the other hand the open ended nature of a question allows students to focus on what exactly is the most important for them.

Students' written comments have not received as much attention as quantitative data from student evaluations. Lots of studies have been done on validity and reliability of quantitative data for course improvement and on relationship between student ratings and student achievements (Cohen, 1981; Feldman, 1989a; Abrami et al., 2007).

Studies on analysis of written comments, that have been published, suggests how written student comments can be organized and analyzed in order to reveal information about aspects of the learning process (Lewis, 2001; Hodges and Stanton, 2007). Most of such studies suggest manual categorization of comments into groups of positive, negative and neutral, or some other kind of grouping, with further investigation of particular factors that reflects students satisfaction or dissatisfaction within each group.

It is quite hard to classify written feedback. Because of it's open-ended nature, the text, that is entered by a student, can range from a few noncritical words such as "cool teacher" to paragraphs with detailed analysis of positive and negative issues of a course, teacher and teaching material. In general, students more often write positive comments, rather then negative, and comments tend to be more general rather than specific (Alhija and Fresko, 2009).

Not much research have been done to investigate the relationship between data obtained from the written comments and data obtained from the quantitative part of evaluations. Improvement of computational power and the development of more sophisticated text mining techniques allows for a more sophisticated

analysis on teacher and course evaluation data.

Studies that have looked into relationship between the quantitative data and the students' written responses suggest that there is a correlation between the quantitative and written feedback from students (Sheehan and DuPrey, 1999), but such examinations are relatively rare.

10.4 Methods

Unstructured data, as students' written feedback, is difficult to process and to analyze. Text mining is the process of deriving information from text, that usually involves the process of structuring the input text, deriving patterns, and finally evaluating and interpreting the output.

Text mining is an interdisciplinary field that draws on information retrieval, data mining, machine learning, statistics, and computational linguistics. It is of importance in scientific disciplines, in which highly specific information is often contained within written text (Manning and Schutze, 1999).

10.4.1 Term-document matrix

A lot of the text mining methods are based on construction of a term-document matrix, high-dimensional and sparse mathematical matrix that describes the frequencies of terms that occur in a collection of documents. There are various ways to determine the value that each entry in the matrix should take, one of them is tf-idf.

Term frequency - inverse document frequency (tf-idf), is a numerical value which reflects importance of a word for a document in a collection of documents. The tf-idf value increases proportionally to the number of times a word appears in the document, but with an offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others (Salton and Buckley, 1988).

Tf-idf is defined as the product of two statistics: term frequency, the number of times that term occurs in a document devided by the total number of words in the document, and inverse document frequency, a measure of whether the term is common or rare across all documents. It is defined by dividing the total

number of documents by the number of documents containing the term, and then taking the logarithm of that ratio.

The tf-idf weight of term t in document d is highest when t occurs many times within a small number of documents, lower when the term occurs fewer times in a document, or occurs in many documents and lowest when the term occurs in almost all documents of a collection.

10.4.2 Key term extraction

Extraction of keyphrases is a natural language processing task for collecting the most meaningful words and phrases from the document. It helps to summarize the content of a document in a list of terms and phrases and thus provides a quick way to find out what the document is about . Automatic keyphrase extraction can be used as a ground for other more sophisticated text-mining methods.

In this study, the Likey keyphrase extraction method (Paukkeri and Honkela, 2010) is used. Likey is an extension of Damerau's relative frequencies method (Damerau, 1993). It is a simple language-independent method (the only language-specific component is a reference corpora). According to the method, a *Likey ratio* (10.1) is assigned to each phrase (Paukkeri et al., 2008).

$$L(p,d) = \frac{rank_d(p)}{rank_r(p)} \tag{10.1}$$

where $rank_d(p)$ is the rank value of phrase p in document d and $rank_r(p)$ is the rank value of phrase p in the reference corpus. The rank values are calculated according to the frequencies of words of the same length n. The ratios are sorted in increasing order and the phrases with the lowest ratios are selected.

10.4.3 Statistical methods

10.4.3.1 Factor analysis

Multivariate data often include a large number of measured variables, and often those variables "overlap" in the sense that groups of them may be dependent. In statistics, factor analysis is one of the most popular methods used to uncover the latent structure of a set of variables. This method helps to reduce

the attribute space from a large number of variables to a smaller number of unobserved (latent) factors.

The most popular form of factor analysis is exploratory factor analysis (EFA), that is used to uncover the underlying structure of a relatively large set of variables. The researcher's a priori assumption is that any indicator may be associated with any factor.

Factor analysis searches for joint variations in response to unobserved latent variables. The observed variables are modeled as linear combinations of the potential factors, plus "error" term. The coefficients in a linear combination are called factor loadings.

Sometimes, the estimated loadings from a factor analysis model can give a large weight on several factors for some of the observed variables, making it difficult to interpret what those factors represent. The varimax rotation is the most commonly used criterion for orthogonal rotation, that helps to simplify the structure and ease interpretation of the resulting factors (Hair et al., 2006).

10.4.3.2 Logistic regression

Logistic regression is a type of regression analysis used in statistics for predicting the outcome of a categorical dependent variable based on one or more usually continuous predictor variables. In cases where the dependent variable consists of more than two categories which can be ordered in a meaningful way, ordered logistic regression should be used.

The relationship between a categorical dependent variable and independent variables is measured, by converting the dependent variable to probability scores. The model only applies to data that meet the proportional odds assumption, that the relationship between any two pairs of outcome groups is statistically the same. The model cannot be consistently estimated using ordinary least squares; it is usually estimated using maximum likelihood (Greene, 2006).

10.5 Data Description

At the Technical University of Denmark (DTU), as in many other universities around the world, students regularly evaluate courses. At DTU students fill final-evaluation web-forms on the university's intranet one week before the final

week of the course. It is not mandatory to fill out the course evaluation. The evaluation form consist of tree parts: Form A contains specific quantitative questions about the course (Table 10.1), Form B contains specific quantitative questions about the teacher and Form C gives the possibility of more qualitative answers divided in 3 groups: What went well?; What did not go so well?; Suggestions for changes.

Table 10.1: Questions in Form A

Id num	Question
A.1.1	I think I am learning a lot in this course
A.1.2	I think the teaching method encourages my active participation
A.1.3	I think the teaching material is good
A.1.4	I think that throughout the course, the teacher has clearly communicated to me where I stand academically
A.1.5	I think the teacher creates good continuity between the different teaching activities
A.1.6	5 points is equivalent to 9 hours per week. I think my performance during the course is
A.1.7	I think the course description's prerequisites are
A.1.8	In general, I think this is a good course

The students rate the quantitative questions on a 5 point Likert scale (Likert, 1932) from 5 to 1, where 5 means that the student strongly agrees with the given statement and 1 means that the student strongly disagrees. For question A.1.6, 5 corresponds to "much less" and 1 to "much more", while for A.1.7, 5 corresponds to "too low" and 1 to "too high". These questions where decoded in such a way that 5 corresponds to best option and 1 corresponds tho the worst.

For this paper data from a Mathematics for Engineers course was analyzed. This is a bachelor 5-ECTS points introductory level course that is available in both spring and fall semesters. The course is well established with almost the same structure over the last 5 years, thus it is large enough to collect a sufficient number of comments to perform text analysis.

Table 12.2 presents the response rates on the course from fall 2007 to spring 2012. The number of students that followed the course during spring semesters is approximatively half of that for fall semesters. The course is mandatory for students who want to enter a Master program at DTU. According to the program the most convenient is to take this course in the fall semester of the second year of education. A part of the spring semester students are those who failed the course in the fall semester. The response rates are lower for spring semesters (33-49%), than for fall semesters (41-62%).

Table 10.2: number of comments

semester	n.s.	n.e.	r.r..	n.p.c.	n.n.c.	n.o.s.
spring 2012	251	85	33,86%	32	28	30
fall 2011	494	239	48,38%	78	60	70
spring 2011	262	93	35,50%	30	41	37
fall 2010	520	212	40,77%	60	46	46
spring 2010	260	101	38,85%	35	25	29
fall 2009	545	337	61,83%	153	91	98
spring 2009	223	73	32,74%	31	22	21
fall 2008	517	290	56,09%	93	71	83
spring 2008	225	111	49,33%	37	21	17
fall 2007	566	326	57,60%	119	58	68
total	3863	1867	48,33%	668	463	499

n.s. - number of students registered for the course
n.e. - number of students participated in evaluation
r.r. - response rate
n.p.c. - number of positive comments
n.n.c - number of negative comments
n.o.s. - number of suggestions for changes

There are more students, who write positive comments than those who write negative. However the average length of the negative comments (35 words) is 10 words larger than the average length of possitive comments(26 words) and suggestions (25 words).

Figure 10.1 shows the average length of positive, negative and suggestion comments.

Figure 10.1: Average length of comments in words

Figure 10.2 shows a change in the average student rating of the course over time. The students satisfaction of the course dropped down by approximately half a point on a Likert scale in spring 2011 for all of the questions except A.1.7. (course prerequisites).

The course is well-established: the curriculum, the book and the structure of

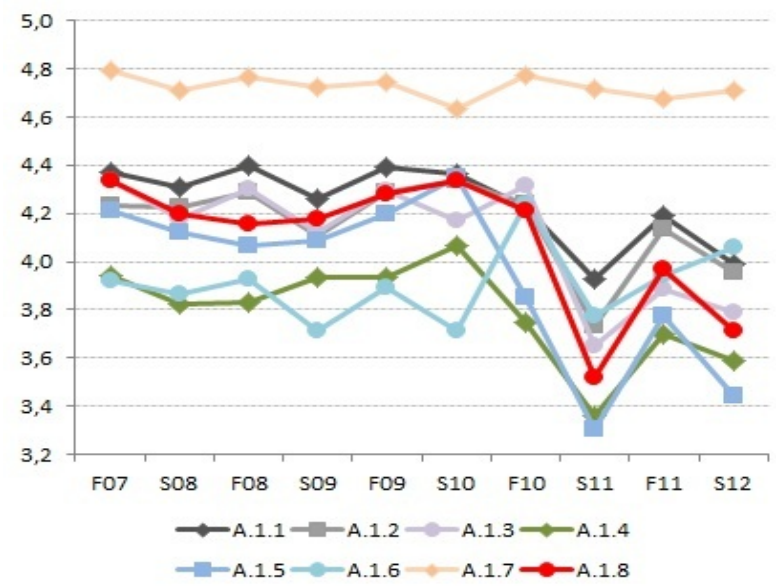


Figure 10.2: Change in average quantitative ratings over time

the course were the same during last years. However one of the main teachers changed in spring 2011. This caused a drop in course evaluation, since the teacher was not experienced in teaching introductory-level courses and had higher expectations to the students. The results of course and teacher evaluations were analyzed and changes in teaching style were made for the next semesters.

The general objectives of the course are to provide participants with tools to solve differential equations and systems of differential equations. Content of the course includes: solution of homogeneous/inhomogeneous differential equations and systems of differential equations, transfer functions, infinite series, power series, Fourier series, applications of infinite series for solving differential equations, the exponential matrix, stability and introduction to nonlinear differential equations. Students also learn how to use Maple to solve problems on the above topics. Some of the above mentioned mathematical issues might be reflected in students comments.

10.6 Results

10.6.1 Term extraction

The length of student comments on the course under consideration ranges from 1 word to 180 words. Even large comments are not long enough to perform keyphrase extraction solely on them. The keyphrase extraction process was done in the following way:

1. All comments for each semester were collected in 3 documents corresponding to the 3 open-ended questions in the questionnaire. It resulted in 10 documents for each type of comments.

2. In order to apply the Likey method, the documents were preprocessed. English comments and punctuation were removed, numbers were replaced with *num* tags and teacher and teaching assistants names with *teachername* and *taname* tags.

3. From each document 50 one-grams (keyphrases that contain just one term - key term) were extracted. These key-terms show the main topics of the students' comments in each semester.

4. Obtained term-lists were stemmed using the Snowball stemmer developed by Porter and irrelevant terms, like slang, were removed.

5. The stemmed term-lists were combined into 3 general term-lists that represent the main topics of comments through the last 5 years.

This procedure resulted in: a positive comments term-list with 142 terms; a negative comments term-list with 199 terms; a term-list of 190 terms representing main topics of suggestions for improvements.

It is not surprising that the negative comments term-list is much longer than the term-list from the positive comments. Students tend to write positive comments that are more general, but in negative comments they tend to write about specific issues they were not satisfied with.

The Danish Europarl corpus, a corpus that consists of the proceedings of the European Parliament from 1996 to present and covers eleven official languages of the European Union (Koehn, 2005), was used as the reference corpus to perform Likey.

Based on these 3 term-lists 3 corresponding term-document matrices where created. Each row correspond to a single comment in the collection of comments over 10 semesters, each column corresponds to a key term and each entry is a tf-idf weight of a key term in the collection of comments. These matrices were used for the further analysis.

10.6.2 Factor analysis

The statistical analysis was done separately for two groups of comments, positive and negative feedbacks. Suggestion comments are expected to correlate a lot with negative comments.

Factor analysis of the term-document matrices was done to reveal the underlying structure of the written feedback from the students. The number of factors, that should be used, is a tricky question, as there is no prior knowledge on the possible number of factors. The Kaiser rule to define the optimal number of factors, that states that the number of factors to be extracted should be equal to the number of factors having variance greater than 1.0, suggests 50 factors for the dataset of positive comments, while randomization method suggests that around 40 factors should be extracted. Another important issue is interpretability of the factors, therefore it was decided to extract 10 factors for each group of comments.

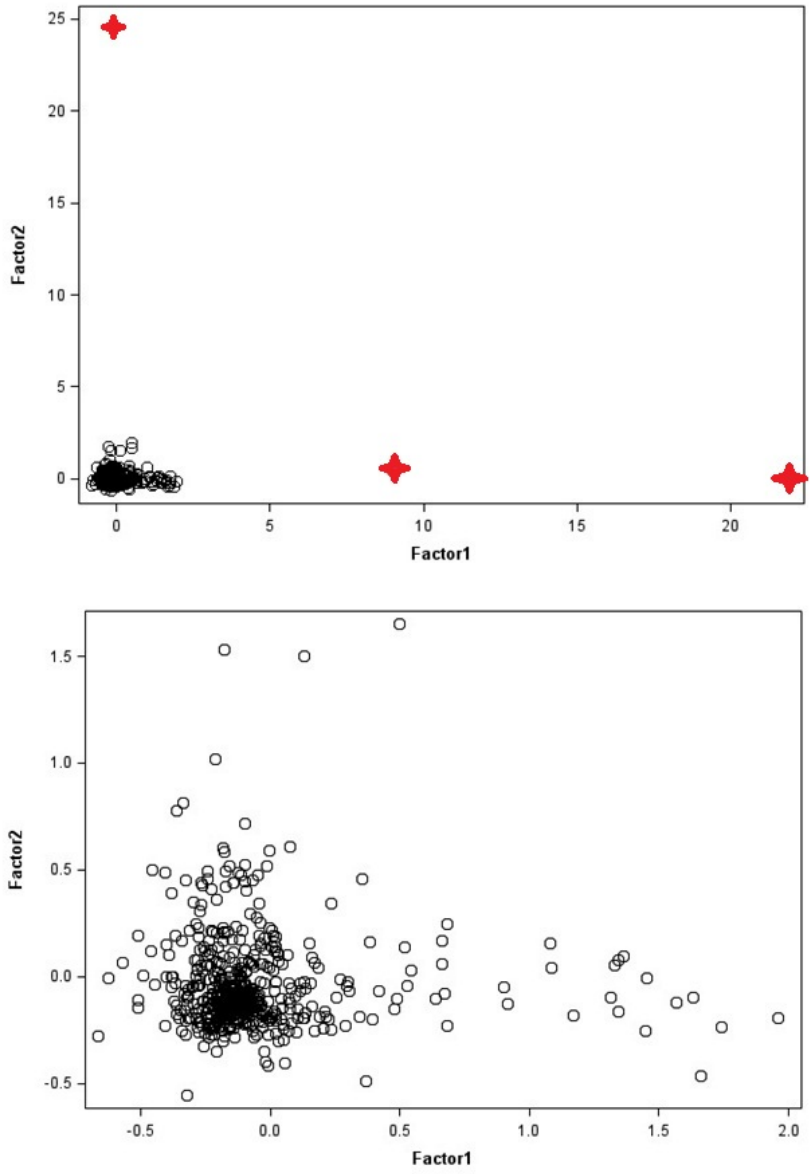
Factor analysis can also be used for outlier detection (Hodge and Austin, 2004). Observations with factor scores, the scores of each case (comment) on each factor (column), greater then 3 in absolute value were considered as outliers.

Figure 10.3 shows the difference of factor scores distribution for the first and the second factor before and after outlier removal for positive comments dataset. At least 3 observations that are different from others.

One of the most illustrative examples of an outlier is a "positive" comment from a student, who had a long break in studying: *"I had a longer break from the studies... when I stopped at the time it was among other things because of this course which frustrated me a lot since... it is nice that this has improved..."*

This comment really differs from the others in the style it is written. Other examples of outliers are comments that mentioned a specific issue that is not mentioned by any other respondents, or comments where a specific issue, for example the "Maple" programming language, is mentioned many times. In total 59 observations were removed from the positive comments data and factor analysis was performed again.

Figure 10.3: Factor1 scores vs. Factor2 scores for positive comments before and after outlier removal



In order to increase interpretability and simplify the factor structure the varimax rotation of the factor reference axes, that aims to have as many zero factor loadings as possible, was done.

Table 10.3 shows the most important variables (factor weight higher than 0.25 in absolute value) in each factor for the positive comments. The presented terms are translated from danish. Terms with are presented.

Extracted factors can be interpreted as:

- Factor1 - *overall course quality in relation to other courses*
- Factor2 - *good teacher qualities.*
- Factor3 - *weekly home assignments* - students were motivated to spend extra hours at home to understand the material.
- Factor4 - *good textbook quality*
- Factor5 - *blackboard teaching performed by lecturer/ presentation of material*
- Factor6 - *"teaching assistant (TA's) communication during exercise classes*
- Factor7 - *weekly question sessions* - question sessions are an extra hours, where students can ask question regarding the course material.
- Factor8 - *teaching during exercise classes.*
- Factor9 - reflects 2 things: *possibility to follow the course twice a week* and *appropriate level of home assignments.*
- Factor10 - *having a good time being a student at the course.*

For the analysis of the negative comments the same outlier removal procedure as for the positive comments was used. It resulted in removing 35 of the negative comments.

Table 10.4 shows the most important terms in each factor, for the negative comments. The factors can be interpreted as follows:

- Factor1 - *Maple as a tool to solve exercises.*
- Factor2 - *English speaking teaching assistants* - students pointed out that it was harder for them to write assignments in English and/or to communicate with English speaking teacher assistants.

Table 10.3: Rotated factor pattern for positive comments

Factor1		Factor2		Factor3		Factor4		Factor5	
keyterm	cor	keyterm	cor	keyterm	cor	keyterm	cor	keyterm	cor
educational	0,60	skilled	0,44	time	0,56	general	0,48	example	0,51
course	0,50	exciting	0,44	assignments	0,47	view	0,45	blackboard	0,40
control	0,41	professional	0,44	additional	0,47	nice	0,45	*teachername*	0,39
DTU	0,36	teacher	0,43	week	0,40	read	0,42	topic	0,39
less	0,36	mathematics	0,39	good	0,36	ok	0,38	go through	0,33
lecturer	0,35	communicate	0,38	home	0,36	course	0,38	really/very	0,32
most	0,31	fun	0,33	idea	0,32	little	0,32	theory	0,29
amount	0,27	*teachername*	0,31	division	0,30	textbook	0,26	statement	0,27
curriculum	0,26	enormous	0,29	understand	0,28	really	0,30	because	0,27
				teachername	-0,30	Maple	0,27	do	0,26
Factor6		Factor7		Factor8		Factor9		Factor10	
keyterm	cor	keyterm	cor	keyterm	cor	keyterm	cor	keyterm	cor
TA	0,63	question session	0,68	lecture	0,36	Monday	0,40	time	0,50
taname	0,59	Tuesday	0,43	really/very	0,35	class	0,36	whole	0,49
good	0,57	week	0,43	exercise	0,33	Thursday	0,34	function/work	0,41
communicate	0,28	teaching material	0,36	good	0,33	great	0,33	students	0,35
very	0,27	pause	0,34	function/work	0,31	amount	-0,27	papershow	0,32
exercises	0,25	course	0,33	material	0,28	home assign.	-0,27	fun	0,25
		fine	0,33	data bar	-0,33	home work	-0,31		
		nice	0,30	nice	-0,38	appropriate	-0,32		
		weekly	0,29	Maple	-0,39	complexity	-0,38		

Table 10.4: Rotated factor pattern for negative comments

Factor1		Factor2		Factor3		Factor4		Factor5	
keyterm	cor	keyterm	cor	keyterm	cor	keyterm	cor	keyterm	cor
Maple	0,70	course	0,61	explain	0,61	teacher	0,36	help	0,60
tool	0,66	englishspeaking	0,49	book	0,56	statement	0,36	teacher	0,59
pity	0,57	think	0,47	stand	0,54	students	0,33	nature	0,57
solve	0,48	TA	0,40	convergence	0,45	better	0,32	often	0,43
possibility	0,41	should	0,39	new	0,41	example	0,31	exercise	0,39
convergence	0,38	understand	0,37	material	0,39	works	0,26	*taname*	0,36
exercise	0,38	mathematical	0,36	fully	0,36	similar	0,26	solution	0,34
whole	0,33	DTU	0,31	example	0,34	subjects	-0,28	more	0,31
give	0,32	really	0,30	poor	0,32	fully	-0,32	hand	0,30
follow	0,30	whole	0,29	read	0,30			difficult	0,29
exam	0,29	Fourier series	0,27	*teachername*	0,28			example	0,27
		used	0,29	lecturing	0,26				
Factor6		Factor7		Factor8		Factor9		Factor10	
keyterm	cor	keyterm	cor	keyterm	cor	keyterm	cor	keyterm	cor
TA	0,49	frustrating	0,48	used	0,49	harder	0,71	room	0,45
grade	0,44	avoid	0,45	difficult	0,45	go through	0,53	campus	0,44
higher	0,42	though	0,44	derivation	0,39	projects	0,51	group work	0,42
difference	0,36	course	0,43	view	0,38	bad	0,43	one	0,38
assignment	0,34	curriculum	0,38	workload	0,36	teaching	0,39	education	0,31
submit	0,33	review	0,38	read	0,34	works	0,39	count	0,31
though	0,32	go through	0,31	project task	0,31	semester	0,38	opposite	0,31
simple	0,27	need	0,31	too much	0,30	away	0,34	annoying	0,29
example	0,27	mathematics	0,31	points	0,30	very	0,32	problem solving	0,29
whole	-0,28	things	0,30	week	0,30	week	0,32	held	0,27
Fourier series	-0,33	start	0,28	good	0,27	assignments	0,28	closer	0,26
understand	-0,34	enormous	0,27	time	0,25			building	0,26
mathematics	-0,39	higher	0,25	very	0,29			mathematics	-0,27

- Factor3 - dissatisfaction with *usage of textbook* - many students argued that examples presented in the class were taken directly from the book.
- Factor4 - *examples to support statements* - some students argue that it was hard to understand some mathematical subjects without examples.
- Factor5 - *not enough TAs for exercise hours*
- Factor6 - *grading of home assignments* - some students complained that TA's grade home assignments differently.
- Factor7 - *frustrating course* - students, that follow the course are very diverse by their background. For some of them the course is really frustrating.
- Factor8 - *project workload* - the course has 2 projects about application of the tools, learned during the course, to the real world problems.
- Factor9 - *last project* - there were complaints that the last project is much harder than the previous.
- Factor10 - *course organization issues: classroom, lecture room and their position on campus.*

10.6.3 Regression analysis

In order to investigate the relationship between the quantitative scores and the qualitative feedback an ordinal logistic regression model was used. Students satisfaction and dissatisfaction points can vary in different semesters, therefore it was decided to investigate which factors were important in which semesters. The number of observations in spring semesters (25-30 comments) is not enough to perform multivariate analysis. Therefore, univariate logistic regression was used for each semester to investigate whether there is an impact of each particular factor on how students rate the course. Question A.1.8 (overall course quality) was used as the response variable.

Table 10.5 shows which positive factors have a significant impact on the way students rate the course. There were no factors, that had a significant impact on the overall course score in spring 2011, the semester when there was a drop in students satisfaction scores (Figure 10.2). However, the next semester four factors: factor2 (teacher qualities), factor3 (weekly home assignment), factor4 (textbook quality) and factor10 (having a good time being a student at the course) had a significant impact on overall rating of the course. It can imply that teachers reacted on results of evaluation and made changes in the course and teaching.

Table 10.5: significance of factors in univariate ordinal logistic regression for question A.1.8 (overall course quality) vs. factors extracted from positive comments

	F07	S08	F08	S09	F09	S10	F10	S11	F11	S12
F1										
F2									**	
F3	**								*	
F4									**	
F5		**								
F6			**	*						
F7										
F8										
F9			*							
F10					*		*		**	

* - significant at 10% significance level
** - significant at 5% significance level

Table 10.6 shows which of the negative factors had significant impact on the way students rate the course. For the spring 2011 semester three negative factors: factor1 (Maple as a tool to solve exercises), factor5 (not enough TAs for exercise hours) and factor9 (last project) had a significant impact. It should be noted that the next semester (fall 2011) none of the negative factors were significant.

Table 10.6: significance of factors in univariate ordinal logistic regression for question A.1.8 (overall course quality) vs. factors extracted from negative comments

	F07	S08	F08	S09	F09	S10	F10	S11	F11	S12
F1								*		
F2					**	*				*
F3										
F4		**								
F5								**		**
F6				*						
F7		**				*				
F8		*			**					
F9		*			**			**		*
F10										

* - significant at 10% significance level
** - significant at 5% significance level

Spring semesters tend to have lower rating than preceding and subsequent fall

semesters (figure 10.2). A similar pattern is observed in the analysis of impact of negative factors on overall course satisfaction: None of the negative factors had a significant impact in fall semesters, except fall 2009. Factor9 (last project) appeared to have a significant impact on overall course satisfaction score in 4 out of 10 semesters. In spring 2011, the new teacher changed the second project completely, but the problem is not only in complexity of the project but also in its placement in the busiest time of the semester, close to the exams period.

Univariate analysis showed that different factors are correlated with the overall course quality score in different semesters. It is not surprising, since each year a new group of students follows the course, teaching assistants are almost always new and teachers can also make changes from semester to semester.

In order to analyze the relationships between the students written feedback and other more specific quantitative evaluations of the course, multivariate logistic regression analysis was used, controlling for year and semester.

Table 10.7 shows which factors, extracted from the positive comments, had a significant impact on the different quantitative evaluation scores of the particular course characteristics (evaluation form A).

Fall semester students, who wrote positive feedback, rated questions A.1.3 (teaching material) and A.1.6 (workload) significantly different from spring semester students.

For the overall measure of satisfaction with the course (A.1.8) only one positive factor - factor5 (presentation of material) had a significant impact, controlling for the semester and year of teaching. There was no factor that had an impact on A.1.4 (feedback from teacher) quantitative score.

For the question A1.1 (learning a lot) 3 factors: factor1 (overall course quality compared to other courses), factor4 (textbook) and factor5 (presentation of material) had a significant impact. Many students appreciate blackboard derivations of theorems and mathematical statements. The book contains illustrative examples, that helps to understand the theory.

Factor1 (overall course quality compared to other courses) together with factor6 (teaching assistant communication) had a significant impact on how students evaluated the teaching method (A.1.2.). It supports the common opinion that teaching assistants play an important role. It is also supported by the fact that factor6 together with factor3 (home assignments) had a significant impact on how students evaluated the teaching method (A.1.3).

There are 3 factors that had a significant effect on how students rate the continu-

Table 10.7: significance of factors in multivariate logistic regressions for course specific questions (Form A) vs. factors extracted from positive comments

Factor	A.1.1	A.1.2	A.1.3	A.1.4	A.1.5	A.1.6	A.1.7	A.1.8
F1	**	**			*			
F2					**		*	
F3			*					
F4	*					**	**	
F5	*							**
F6		**	*					
F7								
F8					**			
F9						**		
F10								
sem(F)			*			**		
y07				**	***		*	*
y08	*							
y09	*		**		**			**
y10					**			
y11	**		***		**			**

* - significant at 10% significance level
** - significant at 5% significance level
*** - significant at 1% significance level

ity between the different teaching activities (A.1.5): factor1 (overall course quality compared to other courses), factor2 (teacher qualities) and factor8 (teaching during exercise classes). The year, the course is performed, also has a significant impact on A.1.5 score. It illustrates the fact that teachers of the course are constantly working on improvements of the teaching methods.

For the evaluation of course workload (A.1.6) high textbook quality (factor4) and complexity of home assignments (factor9) had a significant impact, while prerequisites (A.1.7) teacher qualities (factor2) and high textbook quality (factor4) were important.

Table 10.8 shows which factors, extracted from the negative comments, had a significant impact on the different quantitative scores of course characteristics.

For the overall course quality score (A.1.8), two negative factors appeared to be significant: factor4 (examples to supplement mathematic statements) at 10% significance level and factor7 (frustrating course) at 5% significance level.

Table 10.8: significance of factors in multivariate logistic regressions for course specific questions (FormA) vs. factors extracted from negative comments

Factor	A.1.1	A.1.2	A.1.3	A.1.4	A.1.5	A.1.6	A.1.7	A.1.8
F1								
F2								
F3		**		***	*			
F4						**	**	*
F5			***					
F6	*	*						
F7	**	*	**	*				**
F8						***		
F9		*						
F10		*						
sem(F)								
y07					**			
y08				**				
y09	**	*	*	**	**	*		**
y10			*		**			
y11	*		***		**			**

* - significant at 10% significance level
** - significant at 5% significance level
*** - significant at 1% significance level

Factor1 (Maple) and factor2 (English speaking TAs) appeared to have no significant impact on evaluation of any of the course specific characteristics, when controlling for the time the course were taken.

Factor3 (usage of textbook) is the only negative factor that had a significant (10%) impact on how students evaluate different teaching activities (A.1.5). It also had a strongly significant impact on A.1.4 (feedback from teacher), together with general frustration (factor7). Some of the students complained that examples on the lectures are taken directly from the book, while for others it made reading of the textbook was an easy repetition of the lectures. Question A.1.5 is also rated differently in different years, that illustrates teacher's constant work on improvement of teaching methods.

Factor5 (not enough teaching assistants) had a significant effect only on how students evaluate the teaching method (A.1.3) together with factor7 (frustrating course). In spring 2012 teachers tried to form groups for exercise sessions according to students study lines, to make groups more uniform. But so far it does not have any effect.

For quantitative evaluation scores on question A1.1 (learning a lot) factor6 (grading of home assignments) and factor7 (frustrating course) have a significant impact. Factor8 (project workload) had a significant effect only on how students evaluate the course workload (A.1.6) together with factor4 (examples to support statements).

For the rating of teaching method (A.1.2.) 5 negative factors had a significant effect: factor3 (usage of textbook), factor6 (grading of home assignments), factor7 (frustrating course), factor9 (last project) and factor10 (course organization issues). The last two had an effect only on teaching method evaluation. Evaluation of course prerequisites (A.1.7) is significantly effected only by one negative factor - factor4 (examples to supplement mathematic statements).

To summarize, factors, extracted from the negative comments, had more significant impact on how students quantitatively evaluate different course qualities, than factors extracted from positive comments. The year, the course is taken, also had a significant effect on rating of different course qualities.

10.7 Discussion

The present study is a first step of analysis of relationships between the quantitative and qualitative parts of course evaluation. Further investigations should include analysis of the relationship between the comments and questions the teacher satisfaction questionnaire. It is often reflected in comments, that teachers and teacher assistants play an important role in students satisfaction or dissatisfaction with a course.

Diversity of the students is also an interesting factor that should be taken into account for in future research, in order to investigate whether student specific characteristics such as age, gender, years of education, study line, etc have relationship with the way students evaluate teachers and courses. The diversity of the students backgrounds, ranging from mathematical engineering students, to design and innovation students, may also influence on the high dimensionality of the factorial pattern. Thus it would be of interest to adjust for the student background or to preprocess the data by clustering students.

Regarding the text-mining method used in the analysis, one of the drawbacks is that reference the corpus used in the Likey key phrase extraction is a corpus of very formal language of the European Parliament documentation, while student written comments are usually informal, tend to have some slang phrases and have a lot of course specific technical terms, that get higher weight than other

terms. Another thing is that the Likey method is a purely statistical tool, it does not take synonyms into account. Usage of a more sophisticated main topic extraction tool can improve the results.

10.8 Conclusions

The work makes an analysis of questionnaire data from student-course evaluations from, in particularly the analysis of text from open-ended students comments ant their connection to the quantitative scores.

It was found that factor analysis can help to find comments that are outliers, i.e. really differs from the other in the style it is written or comments about some specific issue that is not mentioned by any other respondent. Furthermore, this method helps to find and summarize the most important points of students satisfaction or dissatisfaction.

It was shown that there is a relationships between some of the factors, extracted from positive and from negative comments, and students' overall satisfaction with the course, and that this relationship changes with the time. It was also shown that different factors have an impact on rating of different course characteristics.

In order to make better responses on students dissatisfaction points and improve courses for the future students, a deeper analysis than just averaging the quantitative data from student evaluation, should be done. Examining the students open-ended feedback from evaluation can help to reveal patterns that can, if properly read, be used to improve courses and teaching quality for future students.

Acknowledgements

Timo Honkela and Mari-Sanna Paukkeri from Department of Informatics and Mathematical Modeling, Aalto University, Helsinki, Finland for helping understanding the text-mining methods.

Chapter 11

Effects of mid-term student evaluations of teaching as measured by end-of-term evaluations: An empirical study of course evaluations

---

Authors: Line H. Clemmensen[1], Tamara Sliusarenko[1], Birgitte Lund Christiansen[2] and Bjarne Kjær Ersbølll[1]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark. [2]LearningLab DTU, Technical University of Denmark.

Paper presented at the 5th International Conference on Computer Supported Education, 5-8 May, 2013, Aachen, Germany

Keywords: End-of-term evaluations: Midterm evaluations: Student evaluation of teaching.

Abstract

Universities have varying policies on how and when to perform student evaluations of courses and teachers. More empirical evidence of the consequences of such policies on quality enhancement of teaching and learning is needed. A study (35 courses at the Technical University of Denmark) was performed to illustrate the effects caused by different handling of mid-term course evaluations on student's satisfaction as measured by end-of-term evaluations. Midterm and end-of-term course evaluations were carried out in all courses. Half of the courses were allowed access to the midterm results. The evaluations generally showed positive improvements over the semester for courses with access, and negative improvements for those without access. Improvements related to: Student learning, student satisfaction, teaching activities, and communication showed statistically significant average differences of 0.1-0.2 points between the two groups. These differences are relatively large compared to the standard deviation of the scores when student effect is removed (approximately 0.7). We conclude that university policies on course evaluations seem to have an impact on the development of the teaching and learning quality as perceived by the students and discuss the findings.

11.1 Introduction

For decades, educational researchers and university teachers have discussed the usefulness of, as well as the best practice for student evaluations of teaching (SET). To a large extent discussions have focused on summative purposes like the use of SETs for personnel decisions as recruitment and promotion (Oliver and Sautter, 2005; McKeachie, 1997; Yao and Grady, 2005). The focus in the present study is the formative aspect, i.e. the use of SETs to improve the quality of teaching and learning.

Much effort has been put into investigating if SETs give valid measurements of teaching effectiveness with students' learning outcome as the generally accepted – though complex to measure – core factor (see metastudies of Wachtel (1998) and Clayson (2009)). Though SETs can be questioned to be the best method for measuring teaching effectiveness (Yao and Grady, 2005), there is a general

agreement that it is the most practical and to some extent valid measure of teaching effectiveness (Wachtel, 1998). Additionally, SETs provide important evidence that can be used for formative purposes (Richardson, 2005).

Studies of the long-term effect of SETs tend to lead to the discouraging conclusion that no general improvement takes place over a period of 3-4 years or more (Kember et al., 2002; Marsh and Hocevar, 1991). However, it is generally found that when the feedback from SETs is supported by other steps, such as consultations with colleagues or staff developers, or by a strategic and systematic approach to quality development at university level, improvements can be found according to the SET results (Penny and Coe, 2004; Edström, 2008).

Some attention has also been directed to the timing of the evaluations – midterm, end-of-term, before or after the exam (Wachtel, 1998). There is some evidence that evaluation results depend on whether they were gathered during the course term or after course completion (Clayson, 2009; Richardson, 2005).

Keeping the formative aim in mind, it is of interest whether midterm evaluations can lead to improvement within the semester to meet the needs of the students in a specific class context (Cook-Sather, 2009). In a meta-analysis of a number of studies comparing midterm and end-of-term SET results, Cohen (1980) concluded that on average the mid-term evaluations had made a modest but significant contribution to the improvement of teaching. His analysis confirms findings from other studies that the positive effect is related to augmentations of the feedback from students – typically consultations with experts in teaching and learning (Richardson, 2005; Penny and Coe, 2004).

In Denmark as in other Nordic countries, the general use of course evaluations has a shorter history. SETs have primarily been introduced for formative purposes as well as an instrument for the institution to monitor and react on student satisfaction in general and on specific issues (e.g. teachers' English proficiency). As an effect of a requirement from 2003, all Danish universities make the outcome of course evaluations public (Andersen et al., 2009). Thus, key results of the existing SET processes are also used to provide information to students prior to course selections.

At the Technical University of Denmark, average ratings of answers to closed questions related to the course in general are published on the university's web site. Ratings of questions related to individual teachers and answers to open questions are not published. The outcome is subject to review in the department study board that may initiate follow-up actions.

As an extensive amount of time and effort is spent on the evaluation processes described, it is of vital interest to examine whether the processes could be im-

proved to generate more quality enhancement. Therefore, the present study provides a basis to consider whether mid-term course evaluations can be used as a supplement to (or partial substitution of) end-of-term evaluations to create an immediate effect on quality of teaching and learning in the ongoing course.

In the study, the student evaluations are treated as a source of information on the teaching and learning process, as perceived by the students, which can be used as a basis for improvements. An experimental setup is designed to address the question: What is the effect of mid-term course evaluations on student's satisfaction with the course as measured by end-of-term evaluations?

The study addresses how general university policies can influence the quality of courses by deciding when to perform student evaluations. Therefore, the course teachers were not obliged to take specific actions based on the midterm evaluations.

The paper is organized as follows. The experimental design is explained in Section 1. Section 2 gives the methods of analysis, and Section 3 the results. Section 4 discusses the findings, and we conclude in Section 5.

11.2 Experimental design

Since 2001 standard student evaluations at the Technical University of Denmark are performed using an online questionnaire posted on "CampusNet" (the university intra-net) as an end-of-term evaluation in the last week of the semester (before the exams and the grades are given). The semesters last thirteen weeks. On one form the student is asked questions related to the course in general (Form A) and on another form questions related to the individual teacher (Form B). The questions can be seen in

Table 12.4. The students rate the questions on a 5 point Likert scale (Likert, 1932) from 5 to 1, where 5 corresponds to the student "strongly agreeing" with the statement and 1 corresponds to the student "strongly disagreeing" with the statement. For questions A.1.6 and A.1.7, a 5 corresponds to "too high" and 1 to "too low". In a sense for these two questions a 3 corresponds to satisfactory and anything else (higher or lower) corresponds to less satisfactory. Therefore the two variables corresponding to A.1.6 and A.1.7 were transformed, namely: 5-abs(2x-6). Then a value of 5 means "satisfactory" and anything less means "less satisfactory". Furthermore, the evaluations contain three open standard questions "What went well – and why?", "What did not go so well – and why?", and "What changes would you suggest for the next time the course is offered?"

Response rates are typically not quite satisfactory (a weighted average of 50%). However, they correspond to the typical response rates for standard course evaluations. The results are anonymous when presented to teachers while they in this study were linked to encrypted keys in order to connect a student's ratings from midterm to end-of-term.

Table 11.1: The evaluation questions

Id no.	Question	Short version (for reference)
A.1.1	I think I am learning a lot in this course	Learning a lot
A.1.2	I think the teaching method encourages my active participation	TM activates
A.1.3	I think the teaching material is good	Material
A.1.4	I think that throughout the course, the teacher has clearly communicated to me where I stand academically	Feedback
A.1.5	I think the teacher creates good continuity between the different teaching activities	TAs continuity
A.1.6	5 points is equivalent to 9 hours per week. I think my performance during the course is	Work load
A.1.7	I think the course description's prerequisites are	Prerequisites
A.1.8	In general, I think this is a good course	General
B.1.1	I think that the teacher gives me a good grasp of the academic content of the course	Good grasp
B.1.2	I think the teacher is good at communicating the subject	Communication
B.1.3	I think the teacher motivates us to actively follow the class	Motivate activity
B.2.1	I think that I generally understand what I am to do in our practical assignments/lab courses/group computation/group work/project work	Instructions
B.2.2	I think the teacher is good at helping me understand the academic content	Understanding
B.2.3	I think the teacher gives me useful feedback on my work	Feedback
B.3.1	I think the teacher's communication skills in English are good	English/English skills

A study was conducted during the fall semester of 2010 and included 35 courses. An extra midterm evaluation was setup for all courses in the 6th week of the semester. The midterm evaluations were identical to the end-of-term evaluations. The end-of-term evaluations were conducted as usual in the 13th week of

the semester. The criteria for choosing courses were that:

1. The expected number of students for the course should be more than 50
2. There should be only one main teacher in the course
3. The course should not be subject to other teaching and learning interventions (which often imply additional evaluations)

The courses were randomly split into two groups: one half where the teacher had access to the results of the midterm evaluations (both ratings and qualitative answers to open questions) and another half where that was not the case (the control group). The courses were split such that equal proportions of courses within each Department were assigned to the two groups. The distribution of responses in the two groups is given in Table 11.2.

Table 11.2: The two groups in the experiment

Access to midterm evaluations	Number of courses	No. of matched responses	Percentage of responses
Yes	17	687	53
No	18	602	46.7

Furthermore the number of students responding at the midterm and final evaluations and the number of students who replied both evaluations are listed. For each question the number of observations can vary slightly caused by students who neglected to respond to one or more questions in a questionnaire.

The majority of the courses were introductory (at Bachelor level), but also a few Master's courses were included. The courses were taken from six different Departments: Chemistry, Mechanics, Electronics, Mathematics, Physics, and Informatics.

No further instructions were made to the teachers on how to utilize the evaluations in their teachings.

11.3 Method

It has been disputed whether, and to what extent, SET ratings are influenced by extraneous factors (Marsh, 1987; Cohen, 1981). In the present study it

is taken into consideration that student evaluations may be biased, e.g. by different individual reactions to the level of grading or varying prior subject interest (Wachtel, 1998; Richardson, 2005), or as a result of systematic factors related to the course such as class size or elective vs. compulsory (McKeachie, 1997; Wachtel, 1998; Aleamoni, 1999). In order to test the differences between midterm and final evaluations as well as differences between with/without access to midterm evaluations while removing factors like students expected grade (Wachtel, 1998; Clayson, 2009) or high/low rated courses, we performed two kinds of tests.

- Paired t-tests where a student from midterm to the final evaluation is a paired observation and we test the null-hypothesis that there is no difference between midterm and final evaluations (Johnson et al., 2011).
- t-tests for the null-hypothesis that there is no difference between having access to the midterm evaluations and not (Johnson et al., 2011). These tests were based on differences in evaluations for the same student in the same course from midterm to end-of-term evaluation in order to remove course, teacher, and individual factors.

In Table 11.2 the number of students who answered both midterm and final evaluations are referred to as the number of matches.

11.4 Results

Pairwise t-tests were conducted for the null-hypothesis that the mean of the midterm evaluations were equal to the mean of the end-of-term evaluations for each question related to either the course or the course teacher. The results are summarized in Table 11.3 and Table 11.4 for the courses where the teacher had access to the midterm evaluation results and those who had not, respectively.

For the courses without access to the midterm evaluations the general trend is that the evaluations are better at midterm than at end-of-term. This is seen as the mean value of the midterm evaluations subtracted from the final evaluations are negative for most questions. In contradiction, the courses with access to the midterm evaluations have a trend towards better evaluations at the end-of-term, i.e. the means of the differences are positive. The question related to the general satisfaction of the course (A.1.8) got significantly lower evaluations at end-of-term when the teacher did not have access to the midterm evaluations ($p = 0.0038$). The question related to the academic feedback throughout the course

Table 11.3: Summary of pairwise t-tests between midterm and end-of-term course and teacher evaluations. For courses without access to the evaluationsn

Final-midterm	Mean difference (std)	p-value	p-value < 0.05
A.1.1 (Learning a lot)	-0.056 (0.96)	0.17	No
A.1.2 (TM activates)	-0.053 (0.98)	0.21	No
A.1.3 (Material)	-0.065 (1.0)	0.13	No
A.1.4 (Feedback)	0.081 (1.1)	0.085	No
A.1.5 (TAs continuity)	-0.075 (1.0)	0.095	No
A.1.6 (Work load)	-0.040 (0.15)	0.53	No
A.1.7 (Prerequisites)	-0.049 (1.2)	0.32	No
A.1.8 (General)	**-0.12 (0.97)**	**0.004**	**Yes**
B.1.1 (Good grasp)	-0.044 (0.86)	0.23	No
B.1.2 (Communication)	-0.066 (0.84)	0.068	No
B.1.3 (Motivate activity)	-0.035 (0.90)	0.36	No
B.2.1 (Instructions)	-0.048 (0.99)	0.33	No
B.2.2 (Understanding)	-0.012 (0.85)	0.78	No
B.2.3 (Feedback)	-0.015 (0.97)	0.76	No
B.3.1 (English)	-0.046 (0.79)	0.54	No

Table 11.4: Summary of pairwise t-tests between midterm and end-of-term course and teacher evaluations. For courses with access to the evaluations

Final-midterm	Mean difference (std)	p-value	p-value < 0.05
A.1.1 (Learning a lot)	**0.09 (0.77)**	**0.004**	**Yes**
A.1.2 (TM activates)	0.048 (0.93)	0.20	No
A.1.3 (Material)	0.019 (0.88)	0.59	No
A.1.4 (Feedback)	**0.18 (1.0)**	**$<$0.001**	**Yes**
A.1.5 (TAs continuity)	0.039 (0.92)	0.29	No
A.1.6 (Work load)	0.058 (1.4)	0.30	No
A.1.7 (Prerequisites)	0.053 (0.93)	0.16	No
A.1.8 (General)	0.039 (0.85)	0.26	No
B.1.1 (Good grasp)	0.020 (0.78)	0.50	No
B.1.2 (Communication)	0.039 (0.74)	0.15	No
B.1.3 (Motivate activity)	0.016 (0.89)	0.64	No
B.2.1 (Instructions)	-0.038 (0.94)	0.36	No
B.2.2 (Understanding)	0 (0.89)	1.0	No
B.2.3 (Feedback)	0.059 (1.0)	0.20	No
B.3.1 (English)	-0.071 (0.73)	0.13	No

(A.1.4) got significantly higher scores at the end-of-term when the teacher had access to the midterm evaluations ($p < 0.0001$). The question related to whether the student felt he/she learned a lot (A.1.1) got significantly higher evaluations at end-of-term when the teacher had access to the midterm evaluations ($p = 0.0040$). The increase or decrease in student evaluations were of average values in the range [-0.12,0.18]), and significant changes were of average absolute values [0.089; 0.18], (A.1.1 with access being the lowest and A.1.4 with access being the highest). The size of the (dis)improvement should be compared with the standard deviations of the differences divided by the squareroot of two (approximately 0.7), which is the standard deviation of the scores where the student effect has been removed.

For the last analysis the midterm evaluations were subtracted from the end-of-term evaluations for each student and each course. The two groups with/without access to midterm evaluations were then compared based on these differences using a two sample t-test for differences between means; the results are summarized in Table 11.5.

Table 11.5: Summary of t-tests of the null-hypothesis that there is no difference in the evaluation differences from midterm to end-of-term between courses with and without access to the midterm evaluations. A folded F-test was used to test if the variances of the two groups were equal. If so, a pooled t-test was used, otherwise the Satterthwaite's test was used to check for equal means.

With-without access	Mean difference	p-value	p-value < 0.05
A.1.1 (Learning a lot)	**0.15**	**0.0045**	**Yes**
A.1.2 (TM activates)	0.10	0.071	No
A.1.3 (Material)	0.084	0.13	No
A.1.4 (Feedback)	0.099	0.11	No
A.1.5 (TAs continuity)	**0.11**	**0.05**	**Yes**
A.1.6 (Work load)	0.098	0.24	No
A.1.7 (Prerequisites)	-0.0037	0.95	No
A.1.8 (General)	**0.16**	**0.0032**	**Yes**
B.1.1 (Good grasp)	0.064	0.18	No
B.1.2 (Communication)	**0.11**	**0.021**	**Yes**
B.1.3 (Motivate activity)	0.051	0.32	No
B.2.1 (Instructions)	0.0095	0.88	No
B.2.2 (Understanding)	0.012	0.84	No
B.2.3 (Feedback)	0.073	0.27	No
B.3.1 (English skills)	-0.025	0.77	No

The general trend is that the courses where the teacher had access to the

midterm evaluation results get a larger improvement in evaluations at the end-of-term than those where the teachers did not have that access (the differences are positive). The only exceptions to this trend are found in two questions regarding factors that cannot be changed during the course (course description of prerequisites (A.1.7) and teacher's English skills (B.3.1)). However, these are not significant. The questions related to the student statements about learning a lot, the continuity of the teaching activities, the general satisfaction with the course, and the teacher's ability to communicate the subject (A.1.1, A.1.5, A.1.8, and B.1.2) had significantly higher increases from midterm to end-of-term when the teachers had access to the midterm evaluations, compared to the courses where the teachers did not have access. Note that the significant differences in means for the questions are of sizes in the range [0.11, 0.16].

According to subsequent interviews (made by phone), the percentage of the courses with access to the midterm evaluations where the teachers say they shared midterm evaluations with students was 53%, and the percentage of courses where the teachers say they made changes according to the midterm evaluations was 53%. The percentage of the courses with access to the midterm evaluations where the teachers say they either shared the evaluations, made changes in the course, or both was 71%.

11.5 Discussion

The results illustrate that students are generally more satisfied with their courses and teachers at end-of-term when midterm evaluations are performed during the course and teachers are informed about the results of the evaluations.

According to the evaluations, students perceive that courses improve when midterm evaluations are performed and the evaluations and the teachers are informed. Though the teachers were not instructed how to react on the results from the mid-term evaluation, it turned out that almost $^3/_4$ of the teachers followed up on the evaluations by sharing the results with their students and/or making changes in the course for the remaining part of the semester. The fact that $^1/_4$ of the teachers acted like the group who were not allowed access to the midterm results could cause the effects to be even smaller than if all teachers acted. The effects are relatively large when compared to the standard deviation of the scores where the student effect has been removed: approximately 0.7.

We expect that the actions upon the midterm evaluations of the $^3/_4$ in many cases have included elaborated student feedback to the teacher, a dialogue about possible improvements, and various interventions in the ongoing teaching and

learning activities, which can explain the increased satisfaction as expressed in the end-of-term evaluation. For this to happen, the teachers should both be motivated and able to make relevant adjustments (Yao and Grady, 2005). The ability to make relevant adjustments will usually increase as a result of participation in teacher training programs that will also encourage teachers to involve both students and peers in teaching development activities. However, less than half of the teachers responsible for the courses in this study have participated in formal University teacher training programs. The proportion of the teachers who have participated in training programs is the same for both groups of courses (35% and 38%, respectively). Therefore, the observed effect of the mid-term evaluation does not seem to be directly dependent of whether the teacher has participated in formal teacher training.

For future work it would be of interest to directly measure the placebo effect of conducting midterm evaluations as opposed to also measuring the effect of real improvement.

From the student comments in the evaluation forms we noticed that there in some courses was a development pointed out. As an example one student writes at midterm that: "A has a bad attitude; Talking down to you when assisting in group work". At end-of-term the student writes: "In the beginning of the course A's attitude was bad – but here in the end I can't put a finger on it". Such a development was found in courses with access to the midterm evaluations and where the instructor said he/she made changes according to the evaluations. This illustrates the usefulness of midterm evaluations when addressing students evaluations within a semester.

In most of the courses the major points of praise and criticism made by the students are reflected both at midterm and end-of-term. Examples are: That the course book is poor, the teaching assistants don't speak Danish, the lecturer is good etc. Thus such points which are easily changed from semester to semester rather than within a semester are raised both from midterm and end-of-term evaluations.

Various studies show that mid-term evaluations may change the attitudes of students towards the teaching and learning process, and their communication with the teacher, especially if the students are involved actively in the process e.g. as consultants for the teachers (Cook-Sather, 2009; Fisher and Miller, 2008; Aultman, 2006; Keutzer, 1993) – and it may even affect the students' subsequent study approaches and achievements (Greenwald and Gillmore, 1997; Richardson, 2005). Such effects may also contribute to the improved end-of-term rating in the cases where teachers with access to the mid-term evaluation results share them with their students.

There is evidence that SETs in general do not lead to improved teaching as perceived by the students (Marsh and Roche, 1993) and one specific study quoted by Wachtel (1998) of faculty reactions to mandatory SETs indicate that only a minority of the teachers report making changes based on the evaluation results.

However, the present study indicates that mid-term evaluations (as opposed to end-of-term evaluations) may provide a valuable basis for adjustments of the teaching and learning in the course being evaluated.

As the course teachers were not obliged to take specific actions based on the mid-term evaluations, the study gives a good illustration of how the university policies can influence the courses by deciding when to perform student evaluations.

It seems to be preferable to conduct midterm evaluations if one is concerned with an improvement of the courses over a semester (as measured by student evaluations). One may argue that both a midterm and an end-of-term evaluation should be conducted. However, it is a general experience that response rates decrease when students are asked to fill in questionnaires more frequently. If this is a concern, it could - based on the results of this study - be suggested to use a midterm evaluation to facilitate improved courses and student satisfaction.

On the other hand, it is widely appreciated that the assessment of students' learning outcome should be aligned with the intended learning outcomes and teaching activities (TLAs) of a course in order to obtain constructive alignment (Biggs and Tang, 2007). Therefore, to obtain student feedback on the entire teaching and learning process, including the alignment of assessment with objectives and TLAs, an end-of-term student evaluation should be performed after the final exams where all assessment tasks have been conducted (Edström, 2008). In this case, teachers can make interventions according to the feedback only for next semester's course. This approach does not facilitate an improvement in courses according to the specific students taking the course a given semester.

Based on the results of the present study it could be suggested to introduce a general midterm evaluation as a standard questionnaire that focuses on the formative aspect, i.e. with a limited number of questions concerning issues related to the teaching and learning process that can be changed during the semester. It should conform to the existing practice of end-of-term evaluations by including open questions and making it possible for the teacher to add questions – e.g. inviting the students to note questions about the course content that can immediately be addressed in the teaching. This can serve as a catalyst for improved communication between students and teacher (Aultman, 2006). As a consequence, the standard end-of-term questionnaire could be reduced and focus on general questions (like A.1.4, A.1.8. and B.1.1, see Table 1) and matters

that are left out in the mid-term evaluation (e.g. teachers proficiency in English, B.3.1). Besides, it could be considered to encourage the teachers to use different kinds of consultations by faculty developers and/or peers to interpret the student feedback (ratings and comments) and discuss relevant measures to take (Penny and Coe, 2004).

The present study considered improvements over one semester as measured by end-of-term student evaluations as opposed to long-term improvements as well as studies including interviews with instructors and students. These limitations were discussed in more detail in the introduction of this paper.

11.6 Conclusions

An empirical study conducting midterm as well as end-of-term student evaluations in 35 courses at the Technical University of Denmark was carried out in the fall of 2010. In half of the courses the teachers were allowed access to the midterm evaluations, and the other half (the control group) was not. The general trend observed was that courses where teachers had access to the midterm evaluations got improved evaluations at end-of-term compared to the midterm evaluations, whereas the control group decreased in ratings. In particular, questions related to the student feeling that he/she learned a lot, a general satisfaction with the course, a good continuity of the teaching activities, and the teacher being good at communicating the subject show statistically significant differences in changes of evaluations from midterm to end-of-semester between the two groups. The changes are of a size 0.1-0.2 which is relatively large compared to the standard deviation of the scores where the student effect is removed of approximately 0.7.

If university leaders are to choose university- or department-wise evaluation strategies, it is worth considering midterm evaluations to facilitate improvements of ongoing courses as measured by student ratings.

Acknowledgements

The authors would like to thank all the teachers and students who participated in the study, the Dean of Undergraduate Studies and Student Affairs Martin Vigild for supporting the project, and LearningLab DTU for assistance in carrying out the study. Furthermore, the authors thank five anonymous reviewers for their valuable comments.

Chapter 12

Respondent Bias in Teacher Evaluations and its Consequences

Authors: Line H. Clemmensen[1], Tamara Sliusarenko[1], Rune Haubo Christensen[1] and Bjarne Kjær Ersbøll[1]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark. Paper submitted to Research in Higher Education

Abstract

Student evaluation of courses and teachers is one of the most common tools to assess and improve teaching effectiveness. One of the problems of these evaluations is that not all the students who participate in a course participate in the evaluation surveys. Such non-participation can lead to biased results of the student evaluations. The aims of the present study are two-fold: first the respondent bias is addressed; secondly the influence of this bias on the evaluation scores is addressed. The study investigates the characteristics of respondents on the basis of the results of course and teacher evaluations. The surveys were conducted at two time points for all involved courses: mid-term and end-of-term. Approximately 60% of the students replied at one or the other time point and 30% replied at both time points leading to increased information of the responders, non-responders, and their behaviours. This was used to examine not only respondent bias, but also the relationship between course and student characteristics and the evaluation scores. It was found that female students, with high grade point average, taking the course for the first time were more likely to participate in the course evaluations. Additionally it was found that the grade obtained on the course was strongly positively correlated with both evaluation participation and the evaluation scores.

12.1 Introduction

Student evaluation of teachers and courses has become the most commonly used measure of teaching effectiveness around the world (Pouder, 2008). The results of such evaluations are usually used by teachers to improve courses (Seldin, 1999; Wright, 2006). University administrations also take the results of teacher and course quality evaluation into account for various personnel decisions such as promotion, tenure, and reappointment. At the same time, student evaluation of teachers (SET) is also one of the most controversial and highly debated measures of course quality. The reliability, validity and various kinds of potential biases of the SETs are often tested (Marsh, 1984). Many researchers have argued that there is no better option that provides the same sort of quantifiable and comparable data on teaching and course quality (Abrami, 2001; McKeachie, 1997).

The traditional way to obtain student evaluation of a course and teacher is to distribute printed questionnaires and survey forms among students at the end of the course, while more modern techniques are based on on-line questionnaires. Regardless of the way SET is conducted, there is a problem of representativeness,

as many students do not participate in the course evaluation. There is some evidence that web-based evaluation methods lead to lower response rates, but that lower response rates do not affect mean evaluation scores (Avery et al., 2006; Donovan et al., 2006; McGourty et al., 2002).Usage of reminder e-mails from instructors or university administration and messages posted on online class discussion boards and forums can significantly increase the evaluation response rates (Norris and Conn, 2005)

The characteristics of the students who actually take part in end-of-course evaluations may have an impact on the results; if students who participate are generally dissatisfied with the course and the instruction, the results will be biased downward, but if students who are satisfied with the course are most likely to evaluate it, the results will be biased upward.

Many universities are switching from paper-based to web-based SETs to decrease costs and facilitate the mode of data collection and analysis, even though they know that response rates can decline. Student's active participation in SET can be critical in the success of such teaching evaluation systems.

When respondents and non-respondents characteristically differ, non-response error in the form of bias can occur. Non-response bias in student evaluations is important to investigate, since the evaluation results seek to be representative and generalizable. As non-response rates increase, the opinions of the responders become less representative of the population. High response rates can reduce the risks of bias (Groves and Peytcheva, 2008).

In this study, in addition to usual end-of-term evaluation, students were asked to evaluate courses and teachers at mid-term as well. The number of students who evaluated at mid-term and end-of-term were close to the usual response rates at DTU. The response rate on each evaluation was around 45%, while around 60% of students participated in either mid-term or end-of-term evaluations. This paper will elaborate more on which students evaluate in the middle of the semester, which evaluate at the end of the semester, and which do not evaluate at all. Additionally, we will seek to answer the question: How do student characteristics influence SET scores?

12.2 Literature

Since the results of student evaluation of teachers and courses are frequently used in faculty personnel decision-making, faculty are concerned about the potential effects of non-response bias. Unlike traditional anonymous paper-based SET,

web-based course evaluation results can be combined with demographic data in order to compare respondents and non-respondents. Although, the actual evaluations remain anonymous, a web-based system provides a possibility to identify students who respond to the survey.

There is evidence suggesting that student ratings are influenced by student personal characteristics like gender, age, race, academic major, expected course grade, GPA, etc.; teacher personal characteristics like teacher's race, gender, rank, experience, etc.; course-specific characteristics like course difficulty, whether the class is compulsory or elective, course size, etc.; academic environment and course organization issues (Martin, 1998; Read et al., 2001). These characteristics can also have an impact on whether or not students actually respond to the evaluation questionnaires.

For example, previous studies showed that high achievers tend to rate their teachers more favorably (Cohen, 1981; Marsh, 2007; McKeachie, 1969). In addition, some authors found that students who earned higher grades on a course and students with higher cumulative GPA are more likely to fill the evaluation forms (Avery et al., 2006; Fidelman, 2007; Porter and Whitcomb, 2005),while the students who are doing poorly in a course are much less likely to submit course evaluations or ratings questionnaires. Hatfield and Coyle (2013) investigated students participation in the evaluation of course and evaluations of faculty. They found several demographic characteristics that correlated with the completion of course and/or faculty evaluations. However, no correlation was found with the obtained course grade and completion of either course or faculty evaluations.

There is some evidence that female students are more likely to evaluate than male students (Thorpe, 2002; Kherfi, 2011; Fidelman, 2007). It was also found that freshmen or new college students are more likely to evaluate a course (Kherfi, 2011).

Student participation in SET can be influenced by student perception of the role SET plays. There is evidence that students are more likely to complete course evaluations if they understand how the evaluations are being used and believe that their opinions have an effect (Gaillard et al., 2006). The negative perception of SET might be less developed among freshmen or new college students. Chen and Hoshower (2003) showed freshmen and seniors have slightly different motivations to participate in the teaching evaluation and several distinct preferences on the uses of the course evaluations.

Characteristics of the course also tended to influence participation (Adams and Umbach, 2012).Submissions of SETs were more likely when the course and the students' major were in the same department. Students who are evaluating

courses that are major requirements in their studies also tend to fill out evaluation forms. They probably think that courses in their own major are more significant than other courses. Thus, students try to participate in course ratings because they think doing so may make a difference.

However, course evaluations are considered a valid and reliable source of information on the quality of teaching (McKeachie, 1997). Many universities now use different strategies and approaches to extract real and honest feedback from their students.

12.3 Methods

12.3.1 Data collection

Students fill out a web- based evaluation form on the university's intranet one week before the final week of the course. The evaluation form consists of three parts:

- 8 questions about the course (rated on a 5 point Likert scale).
- 6 questions about each teacher and/or teaching assistant (rated on a 5 point Likert scale).
- Three open-ended questions where students can write their feedback.

An extra mid-term evaluation, identical to the end-of-term evaluations, was set up for 35 selected courses in the 6th week of the fall 2010 semester. The criteria for choosing courses were that:

- The expected number of students for the course should be more than 50.
- There should be only one main teacher on the course, who performs a majority of the lectures.
- The course should not be subject to other teaching and learning interventions (which often imply additional evaluations)

The courses were split such that equal proportions of courses within each department were randomly assigned to the two groups. One half where the teacher had access to the results of the mid-term evaluations (both ratings and qualitative answers to open questions) and another half where this was not the case (the control group).

Table 12.1 provides the list of the courses in the experiment, the number of students registered on each course, as well as the response rates in the mid-term and end-of-term evaluations. There is an issue of completeness of evaluation forms. Some students may not have answered all questions leaving the evaluation incomplete. We included all students that answered at least one question in the questionnaire in this study.

Table 12.1: The Response Rates for the courses under investigation

Course name	Teacher access to the mid-term	# of students registered	Mid-term response rate, %	End-of-term response rate,%	Mid-term & end-of-term response rate, %
Chemistry					
Inorganic Chemistry	no	52	46,2	61,5	38,5
Physical Chemistry	no	67	58,2	44,8	41,8
Organic Chemistry	yes	105	51,4	49,5	39,0
Electronics					
Electric circuits 1	yes	117	47,9	41,0	29,1
Electronics	no	74	32,4	18,9	10,8
Engineering Electromagnetics	no	60	50,0	61,7	40,0
Informatics					
Embedded systems	no	73	63,0	50,7	41,1
Digital Electronics 1	no	78	55,1	43,6	33,3
Data Security	yes	88	44,3	35,2	30,7
Software Development of Web Services	no	73	43,8	28,8	17,8
Introductory Programming	yes	62	37,1	25,8	14,5
Development methods for IT-Systems	yes	90	35,6	18,9	11,1
Probability and Statistics	no	239	31,8	45,2	21,8
Windows Programming using C# and .Net	no	81	43,2	35,8	17,3
Programming in C++	yes	76	48,7	32,9	28,9
Introduction to Statistics	yes	319	39,5	52,4	31,7
Probability theory	yes	96	29,2	25,0	17,7
Multivariate Statistics	yes	80	52,5	45,0	36,3
Introduction to Numerical Algorithms	no	62	37,1	30,6	22,6
Optimization and Data Fitting	yes	77	55,8	51,9	37,7

Continued on next page

Table 12.1 – *Continued from previous page*

Course name	Teacher access to the mid-term	# of students registered	Mid-term response rate, %	End-of-term response rate,%	Mid-term & end-of-term response rate, %
Introductory Programming with Matlab	no	166	44,6	50,6	31,3
Web 2.0 and mobile interaction	yes	90	44,4	27,8	20,0
Mathematics					
Advanced Engineering Mathematics 2	yes	520	50,0	46,3	32,1
Calculus and algebra 1	no	455	50,8	33,6	24,4
Calculus and algebra 2	no	214	36,0	29,9	19,6
Geometric Operations in Plane and Space	yes	88	56,8	40,9	33,0
Mechanics					
Hydrodynamics	no	57	52,6	54,4	45,6
Plate and Shell Structures	yes	45	60,0	55,6	44,4
Statics	yes	109	60,6	36,7	29,4
Computational Fluid Dynamics	no	50	76,0	64,0	56,0
Fracture Mechanics	yes	48	66,7	58,3	56,3
Mechanics	no	94	52,1	31,9	23,4
Physics*					
Physics 1	no	216	47,2	45,4	33,3
Physics 1	yes	294	42,2	36,4	25,5
Physics 1	no	72	34,7	37,5	27,8

* - Physics 1 course has different versions, depending on a student major. All 3 versions are lectured by different teachers.

12.3.2 Statistical methods

12.3.2.1 Logistic regression

Logistic regression is used in statistics for predicting the outcome of a categorical dependent variable. If the dependent variable consists of more than two categories which can be ordered in a meaningful way, ordered logistic regression should be used. The model is usually estimated using maximum likelihood (Greene, 2006).

The relationship between a categorical dependent variable and independent variables is measured by converting the dependent variable into probability scores. The model only applies to data that meet the proportional odds assumption.

Logistic regression models the log odds of participation in SET as a linear combination of the predictor variables. For the continuous variables the estimated coefficients can be interpreted in the following way: for one unit increase in the variable, the log odds of submission increases/decreases by the estimated coefficient.

12.3.2.2 Variable selection

Stepwise selection is the most commonly used method to select the most important variables out of the large numbers of potential independent variables in a model. At each step variables are added or removed from the model based on the significance of their estimated coefficients.

12.3.2.3 Imputation

The problem of missing data exists in many data sets. Common reasons for missing data in surveys include refusal to answer, insufficient knowledge, and loss of contact. Imputation is the process of replacing missing data with substituted values. There are different approaches to deal with missing values in a data set.

Multiple imputation, developed by Rubin (Rubin, 1987), is an attractive approach for analysing incomplete data. The method starts by running stochastic regression on the same data set multiple times and the imputed data sets are saved for later analysis. Then each missing value is replaced with a set of plausible values that represent the uncertainty about the right value to impute. The multiple-imputed data sets are then analysed by using standard procedures for complete data and combining the results from these analyses.

12.4 Data

One of the advantages of the web-based student evaluation of teaching is that the results of the evaluations can be combined with demographic data, keeping

students' anonymity. Additionally, it was possible to combine the course evaluation from each student with the student's grades, and the corresponding course characteristics..

12.4.1 Students and courses characteristics

Among course characteristics, course size, experience of the teacher on the course, course workload (ECTS points), course level, and course language are available. Student-specific characteristics that are available include student's age, gender, nationality, year of entering the university, study program, study line, type of entrance exam, high school, high school GPA, university GPA, and grades in mathematics, physics and chemistry from high school.

The study concentrates only on bachelor, diploma and master's students, therefore guest students, open university students, students from the ERASMUS master's program and PhD students were not considered. These students only represented 3.5% of the initial sample, moreover some student-specific variables like study line or GPA were missing for these students.

The final sample consists of 4211 observations representing information on 2574 individual students. Table 12.2 presents the different course-specific characteristics.

Table 12.2: Descriptive statistics of course specific variables

	# of courses	% of courses	# of students	% of students
Courses	35		4211	
Department				
Chemistry	3	8,6	220	5,2
Electronics	3	8,6	243	5,8
Informatics	16	45,7	1547	36,7
Mathematics	4	11,4	1216	28,9
Mechanics	6	17,1	362	8,6
Physics	3	8,6	623	14,8
Course size				
<100	24	68,6	479	11,4
100-200	4	11,4	1004	23,8
200-300	4	11,4	1219	28,9
300+	3	8,6	1509	35,8
Course language				
Danish	26	74,3	3733	88,6
English	9	25,7	478	11,4

Continued on next page

Table 12.2 – *Continued from previous page*

	# of courses	% of courses	# of students	% of students
Course workload (ECTS points)				
5	26	74,3	3231	76,7
10	9	25,7	980	23,3
Course level				
Bachelor	18	51,4	2413	57,3
Diploma	10	28,6	1438	34,1
Master	7	20,0	360	8,5
Teacher know results of the mid-term evaluation				
no	18	51,4	2072	49,2
yes	17	48,6	2139	50,8
Visiting lecturers				
no	25	71,4	2508	59,6
yes	10	28,6	1703	40,4
Semesters per year				
1	22	62,9	1431	34,0
2	13	37,1	2780	66,0
Time of the day				
Evening	11	31,4	1036	24,6
Morning	24	68,6	3175	75,4
Day of the week				
Monday	8	22,9	1556	37,0
Tuesday	9	25,7	888	21,1
Wednesday	7	20,0	605	14,4
Thursday	3	8,6	386	9,2
Friday	8	22,9	776	18,4
Course twice a week				
no	27	77,1	2724	64,7
yes	8	22,9	1487	35,3

Table 12.3 presents the descriptive statistics of the various student characteristics. Student study lines were used to construct the student department variable and also to find out whether the particular course was mandatory or elective.

Table 12.3: Students descriptive statistics

	number	%		number	%
Students	4211		Students	4211	
Gender			**Student department**		
Female	898	21,3	Chemistry	400	9,5

Continued on next page

Table 12.3 – *Continued from previous page*

	number	%		number	%
Male	3313	78,7	Civil Engineering	785	18,6
Age group			Electrical Eng.	764	18,1
<20	225	5,3	Environment	94	2,2
20-25	3350	79,6	Food	14	0,3
25-30	422	10,0	Fotonics	78	1,9
30+	214	5,1	Informatics	893	21,2
Student level			Mathematics**	270	6,4
Bachelor	2061	48,9	Management	40	0,9
Diplom	1681	39,9	Mechanical Eng.	556	13,2
Master	469	11,1	Physics	61	1,4
Years at DTU			System Biology	211	5,0
First year	553	13,1	Transport	34	0,8
Second year	1089	25,9	Unknown	11	0,3
Third year	638	15,2	**Student gymnasium region**		
More then 3 years	1789	42,5	Capital Region	2532	60,1
Unknown	142	3,4	Central Denmark	198	4,7
First time at course			North Denmark	74	1,8
no	479	11,4	Southern Denmark	377	9,0
yes	3626	86,1	Zealand	731	17,4
Unknown	106	2,5	Outside Denmark	42	1,0
Entrance exam type			Unknown	257	6,1
ordinary exam	2792	66,3	**Nationality**		
technical exam	432	10,3	Danish	3854	91,5
other exam	920	21,8	Foreign	357	8,5
Unknown	67	1,6	**DTU GPA**		
Pre DTU GPA			0-3	39	0,9
0-3	46	1,1	3-4	116	2,8
3-4	128	3,0	4-5	317	7,5
4-5	210	5,0	5-6	407	9,7
5-6	425	10,1	6-7	449	10,7
6-7	541	12,8	7-8	475	11,3
7-8	574	13,6	8-9	438	10,4
8-9	676	16,1	9-10	245	5,8
9-10	501	11,9	10-11	227	5,4
10-11	342	8,1	11-12	90	2,1
11-12	290	6,9	Unknown	1408	33,4
Unknown	478	11,4	**Student from course department**		
Course is mandatory			no	2931	69,6
no	1576	37,4	yes	1280	30,4
yes	2635	62,6			

* - real age will be used in the modeling

Continued on next page

Table 12.3 – *Continued from previous page*

	number	%	number	%

** - programs are shared between DTU Mathematics and DTU Informatics

Some of the variables have missing values for various reasons. The most problematic variable is GPA from this university, which has missing values for one-third of the sample. Part of these missing values comes from first-year students who do not have university GPA in their first semester of studying. In order not to miss a significant part of the sample during model estimation, the university GPA and entrance GPA variables have been imputed using the multiple imputation technique.

12.4.2 Evaluation survey

Table 12.4 presents the questions from the first part of the quantitative part of the survey used at the University (questions about the course). The students rate the questions on a 5 point Likert scale (Likert, 1932) from 5 to 1, where 5 corresponds to the student "strongly agreeing" with the statement and 1 corresponds to the student "strongly disagreeing" with the statement. For questions A.1.6 and A.1.7, a 5 corresponds to "too high" and 1 to "too low". In a sense, for these two questions, a 3 corresponds to satisfactory and anything else (higher or lower) corresponds to less satisfactory. Therefore the two variables corresponding to questions A.1.6 and A.1.7 were transformed in such a way that a value of 3 becomes 5 “satisfactory”, 2 and 4 becomes 3 "less satisfactory", 1 and 5 becomes 1 “least satisfactory”.

12.5 Results

12.5.1 Sample

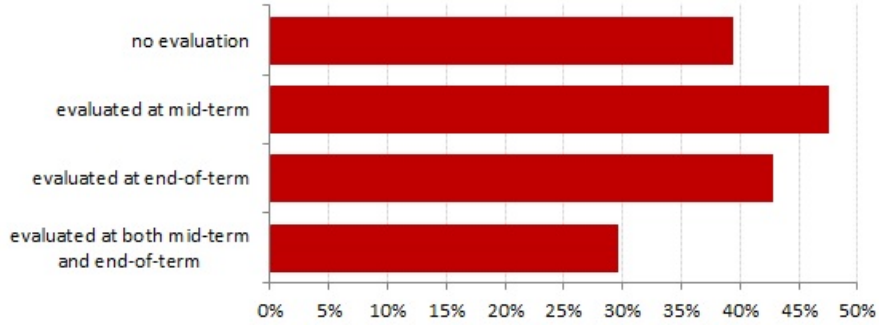
The average response rates on mid-term and end-of-term course evaluations for the courses in the experiment do not exceed 50%. Figure 12.1shows the percentage of students in the sample who participated in mid-term, end-of-term, both or none of the course evaluations.

Students who did not submit mid-term or end-of-term evaluation forms are, on average, half a year older and spend more years at the university, have lower

Table 12.4: Questions of Form A: evaluation of course quality

Id no.	Question	Short version
A.1.1	I think I am learning a lot in this course	Learning a lot
A.1.2	I think the teaching method encourages my active participation	TM activates
A.1.3	I think the teaching material is good	Material
A.1.4	I think that throughout the course, the teacher has clearly communicated to me where I stand academically	Feedback
A.1.5	I think the teacher creates good continuity between the different teaching activities	TAs continuity
A.1.6	5 points is equivalent to 9 hours per week. I think my performance during the course is	Work load
A.1.7	I think the course description's prerequisites are	Prerequisites
A.1.8	In general, I think this is a good course	General

Figure 12.1: Percentage of students who participated in SETs



GPA, lower entrance grades and obtained lower course grades. This may imply that better students are more willing to evaluate the course and the teacher. Therefore, the whole evaluation may be biased, because it mainly represents the opinions of good students.

The dataset has a problem of missing values mainly in student-specific characteristics (see Table 12.3). Only 2365 out of 4211 observations were without missing values. The multiple imputation technique was used to impute the missing values for university GPA and entrance GPA variables.

12.5.2 What effects student participation in SETs

Logistic regression with stepwise selection was used to fit the two models: one for the students' SET participation in the mid-term evaluation, another for the end-of-term evaluation.

Table 12.5.3 presents the variables that significantly improved the model fit (under 5% significance level) selected by the stepwise procedure, together with the estimated model coefficients and odds ratios.

The interpretation of the estimated coefficients for continuous variables is the following: one unit increase in GPA increases the log odds of participation in SET by 0.06 for the mid-term evaluation and by 0.09 for the end-of-term evaluation. Similarly, for one unit increase in entrance GPA, the log odds of submission increases by 0.11 for the mid-term and by 0.08 for the end-of-term evaluation.

For the categorical variable, it is easier to interpret the log odds ratios, which can be interpreted as the multiplicative change in the odds for a one-unit change (for continuous variables) or for category change vs. the reference category (for categorical variables). For example, the probability of submitting the SET at the mid-term/end-of-term evaluation, for those students who repeat the course were by a factor 0.36/0.53 lower than for those following the course for the first time. Female students were more likely to participate in both mid-term and end-of-term evaluations by a factor of 1.40 and 1.65, respectively, than male students.

Course-specific characteristics had different effects on SET participation at the two different time points. The course workload, course teaching language, visiting lecturers and whether the student is from the same department as the course had no impact on whether students participate in evaluation at the middle of the semester, but had a significant effect on end-of-term SET participation. In contrast, whether a course runs twice a year or twice a week had a significant impact on participation in the mid-term SET, but not in the end-of-term SET.

In addition, the coefficients of some of the variables that had a significant effect on SET participation in both the mid-term and the end-of-term SET had different signs. For example, for the course department variable, students that follow the courses taught by all other departments vs. the department of Physics were less likely to evaluate at the mid-term, but more likely to evaluate at the end of term. Moreover, for some of the significant variables, coefficients had different magnitudes in the two models. For obtained grade, the estimated log odds were higher for the end-of-term model than for the mid-term model.

Table 12.5: Results of Logistic Regression: the significant predictors of course evaluation survey response in the middle of the semester and at the end of the semester.

	Effect	mid-term participation			final term participation		
		Coef.	p-value	odds ratio	Coef.	p-value	odds ratio
Intercept		-1,97	*		-4,55	***	
Course department (ref=physics)	Chemistry	-0,39		0,68	2,31	***	10,07
	Electronics	-0,45		0,64	1,76	***	5,83
	Informatics	-0,73	*	0,48	1,27	***	3,55
	Mathematics	-0,87		0,42	0,88	**	2,42
	Mechanics	-0,05		0,95	2,12	***	8,29
Course workload	5 ECTS				-0,51	**	0,81
Course language	Danish				1,06	**	2,90
Course level (ref=master)	bachelor	-0,58	*	0,56	-1,39	***	0,25
	diplom	-0,44		0,64	-1,98	***	0,14
Semesters per year	1	0,44	*	1,55			
Course size (ref=<100)	100-200	-0,06		0,94	-0,17		0,85
	200-300	-0,22		0,80	0,43	***	1,53
	300+	0,32		1,38	0,81	***	2,25
Course twice a week	no	-0,79	**	0,46			
Visit lecturers	no				0,48	***	1,61
DTU GPA		0,06	**	1,07	0,09	***	1,09
Entrance GPA		0,11	***	1,12	0,08	***	1,08
Entrance exam (ref=tech. exam)	ordinary exam	-0,13		0,88	-0,28	**	0,76
	other exam	0,97	*	2,63	-0,35		0,70

Continued on next page

Table 12.5 – *Continued from previous page*

	Effect	mid-term participation			final term participation		
		Coef.	*p*-value	odds ratio	Coef.	*p*-value	odds ratio
First time with the course	no	-1,03	***	0,36	-0,63	***	0,53
Student is from course department	no				0,26	*	1,30
Gender	female	0,34	***	1,40	0,5	***	1,65
Obtained grade (ref='no show')	not passed (-3, 0)	0,93	***	2,53	1,51	***	4,53
	passed (2, 4)	1,08	***	2,96	1,67	***	5,29
	7	1,22	***	3,37	1,99	***	7,34
	10	1,49	***	4,43	2,22	***	9,17
	12	1,78	***	5,91	2,26	***	9,58

Significance levels: * $p <$,05; ** $p <$,01; *** $p <$,001;

Continued on next page

12.5.3 Relationships between student and course characteristics and SET scores

In addition to the analysis of whether student and course characteristics affect participation in a student evaluation, the data gives the possibility to investigate whether the same characteristics have an impact on the SET scores.

In order to perform the analysis, all the students' scores in both mid-term and end-of-term evaluations were pooled into one dataset. Table 12.5.3 shows that course-specific and student-specific variables had significant impact on SET scores for different aspects of course evaluation. The results were obtained using ordered logistic regression for each question in the evaluation questionnaire.

Different student-specific and course-specific variables had a significant impact on how students rate different course aspects. Almost all available variables were significant for one or another question of course evaluation. Only the variable 'first time on the course' had no effect on the evaluation scores of the different course aspects, however the variable was highly significant for the SET participation models.

The GPA from this university had an influence on the SET participation, but for the SET scores it only had a significant effect on how students answer question A.1.7 (Prerequisites). The effect is negative, meaning that the higher the GPA, the lower the score the course gets on A.1.7, which in turn means that students found the course prerequisites too high or too low.

Five variables: course department, course size, course weekday, student gender and obtained grade, were found to be significant in all or almost all the models of how students evaluate different course aspects. Males tended to give lower scores on evaluations than females in all the questions, when other variables are held constant. The pattern for course department variable was very similar for all the questions in the SET. The students from courses from the departments of Mathematics, Informatics, Electronics, Mechanics and Chemistry were less likely to give higher score to different course aspects than the students in courses from the department of Physics.

Regarding course size, it appeared that courses with 200-300 registered students got significantly lower SET scores compared to courses with 50-100 registered students for all survey questions except A.1.7. Estimates for other course sizes, namely 100-200 and more than 300 registered students, were not significantly different from the reference group of courses with 50-100 students.

Apparently, the day of the week on which the course is taught also had an

Table 12.6: Logistic regression results: Significance of course-specific and student-specific characteristics that has an effect on different aspects of the course.

	A.1.1 learning lot	A.1.2 TM activities	A.1.3 Material	A.1.4 Feedback	A.1.5 TAs continuity	A.1.6 Work load	A.1.7 Prerequisites	A.1.8 General
Course specific characteristics								
Course department	***	***	***	***	***	***	**	***
Workload	***		**		***	***	***	***
Teacher know the midterm results	***	***	***	**	**			***
Course language				**			***	***
Course level	***	***	**		***	**	***	
Semester per year	***				**			
Course size	***	***	***	***	***	***	***	***
Teacher experience	***	***			**			***
Time day	***				**			
Twice a week	***		***	***	***		***	***
Visiting lecturer	***		***		**	**		***
Weekday	***	***	***	***	***	**	***	***
Students specific characteristics								
Student age						***		
DTU GPA							***	
Entrance GPA		**	**	***				
Entrance exam type							**	**
First time at the course								

Continued on next page

Table 12.6 – *Continued from previous page*

	A.1.1 learning lot	A.1.2 TM activities	A.1.3 Material	A.1.4 Feedback	A.1.5 TAs continuity	A.1.6 Work load	A.1.7 Prerequisites	A.1.8 General
Student from course dep			**	***				
Gender	***	***	***	***	***	***	***	***
Obtained grade	***	***	***	***	***		***	***
Course is mandatory	**		**		***	**		***
Nationality			***					
Student program	***		*		***			
Study time				***				
Evaluation period			***	***		**		

significance levels: ** $p <$,05; *** $p <$,01; * $p <$,001;

Continued on next page

impact on how students rate all the aspects of the course. Courses performed on Mondays were less likely to receive higher scores than courses performed on Wednesdays, while courses performed on Friday were more likely to receive higher scores. At DTU the block structure of the courses means that Wednesday courses would be 10 ECTS whole day courses, while other courses are either 5 ECTS courses or split into several modules per week. The student's obtained course grade was correlated with both students' SET scores and students' SET participation. The findings here suggest that students with higher obtained course grades were more likely to give higher evaluation scores to the different aspects of the course. Since students evaluate the courses before the final exam and before they get their final grade, the evaluations were not affected by student performance on the final exam or its complexity.

The fact whether the teacher of the course had an access to the results of mid-term evaluations had significant effect on all SET scores except the scores of questions A.1.6 and A.1.7. If the teacher did not have access to the mid-term evaluation results, the course tended to receive lower end-of-term evaluation scores compared with a course where teacher had an access to the results of mid-term evaluations.

12.6 Discussions

12.6.1 SET participation

The mid-term evaluation was conducted in addition to regular end-of-term evaluations in the fall semester 2010 in order to check whether the these evaluations can lead to improvement within the semester to meet the needs of the students in a current class, not just future students. It is a general experience that response rates decline when students are asked to participate in surveys more frequently. The average response rate on the mid-term evaluation was 47.9%, while for the end-of-term it declined to 41.4%. Around 30% participated in both evaluations, while 60% of students participated in either mid-term or end-of-term evaluation (Figure 12.1).

The results illustrated that different course-specific and student-specific variables had an effect on whether a student participated in the student evaluation of teaching in the middle or at the end of a semester. Some of the variables had similar effects, while others had opposite effects, different magnitude or no effect at one of the time points.

It can be concluded that the general profile of students that were more likely to participate in SET in both time points were female students with high GPAs (both current and from the high school) taking the course for the first time. These characteristics were found to have similar effects on SET participation in previous studies (Avery et al., 2006; Fidelman, 2007). However, none of the previous studies investigated student evaluation participation at two time points. Some of the studies also found that students' age, students' study time and whether or not the course is compulsory had an impact on SET participation. However, none of the above mentioned variables were significant for the sample under investigation.

Whether the teacher on the course was informed about the results of the evaluations had no effect on students' med-term and end-of term SET participation. The students of the courses under experiment did not know if the teacher had access to the mid-term evaluation results. For the end-of term SET participation there are two possible effects that might offset each other. Students from the courses where the teacher had access to the results could participate in end-of-term SET in order to appreciate the observed changes in the second part of the course, while students from the second group of courses could participate in SET in response to unobserved changes, arguing that their opinion was not heard.

Some variables were significant for mid-term SET participation but not for the end-of-term SET participation and vice versa. For example, whether the course was performed by the same department as the students' major had no significant effect on whether students participated in the mid-term evaluation, but had a positive effect on participation in the end-of-term evaluation. Students might feel more responsibility for courses and teaching quality for departments they belong to and participate in the regular end-of-term evaluation, but they are not willing to participate in an extra evaluation.

Concerning the obtained course grade variable, the higher the grade the student obtained in the course, the higher was the probability that the student had participated in SET. However, the magnitude of estimated coefficients was higher for the end-of-term evaluation. Looking deeper into the results, the reference category for the obtained grade variable was "no show for the exam". Most of students who decided to drop out of the course did not officially unregister from the exam; they just stopped attending the lectures. Some students decided to drop out after just a few first lectures, but others decided in the middle of the course or right before the exam. Therefore the fact that a student participated in the exam already makes this student more likely to participate in SET.

Under the current course evaluation and course registration systems set up it is impossible to distinguish between students who dropped out of the course at

the beginning of the semester or at the end. However, Crews and Curtis (2011) in their suggestions for online course evaluation systems noted the importance of ensuring that when students withdraw from a course they are dropped from the evaluation system.

Student survey participation can be related to academic discipline. There are two natural effects to consider. One is that an academic environment encourages students to value SET completion. Another is that, if students are taking a course in their department, they feel the need to support the environment and invest in their academic departments by evaluating the courses. However, students in majors outside of that department may complete the evaluation depending on whether the topic of the course was applicable or interesting to them.

Regarding course departments, the reference category was "the department of Physics". The three courses, which were selected for the experiment, consist of fall and winter 13-weeks semesters, but with different topics and another teacher in the second semester of the course. For these students the final evaluation was like a second mid-term evaluation, therefore they were probably much more engaged in the SET after 6 weeks of the course, but on the 13th week evaluation students from courses from other departments have a higher probability to evaluate.

Overall the student-specific characteristics that had an impact on students' participation are the same for the mid-term and end-of-term SETs, but the course-specific characteristics had different effects in mid-term and end-of-term. There is evidence that students are more likely to complete course evaluations if they understand how the evaluations are being used (Gaillard et al., 2006). It is clear, that the mid-term evaluation attracted some students that typically are "non-respondents".

It can be concluded, that the results of course evaluations represent mostly the opinion of high-achievers. We believe university administration and teachers should be aware of this fact, while making the course adjustments and personnel decisions. Knowing characteristics of non-respondents allows survey administrators to direct their efforts to reach these students and reduce non-response error in the future, and consequently get more truthful ratings of the course and teaching quality.

12.6.2 SET scores

Previous studies of student course evaluations have found that students with higher grades typically award teachers with higher SET scores (Johnson, 2003b; Weinberg et al., 2007). According to the results of this study, the students with higher grades are more likely to respond on SETs and to reward the course with higher SET Scores.

There are two natural hypotheses on the positive correlation between obtained grades and course evaluation scores. One suggests that professors might lower their grading policy and make the course easier and more fun in order to get better evaluations, while another hypothesis is that the good teachers are able to motivate students to study hard, and consequently get better grades. At our university, students submit the evaluation before getting their final grade, so the SET scores can only be affected by the students' grade expectations and grades on home assignments or projects obtained during the semester, which are sometimes also part of the final grade. Some courses are subject to external censors for the final examinations in order to give more objective grades, but this does not hold for all courses.

It is commonly accepted that SET scores are influenced by the expected grades. Previous research by Nowell (2007) found that as expected grade increases, SET scores increase, and as the historical GPA increases, the SET scores decline. If student's expectations are higher than their historical GPA, students may reward teachers with higher SET scores. However, in this campus for the courses under experiment it was found that students with higher GPA do not give the courses higher SET scores. In this context, it should be noted that there is some correlation (0.54) between GPA and obtained graded, but thus also a distinction.

Apart from the obtained grade, according to the results, many other factors had an effect on how students rate courses. Some were important for all of the course aspects (course department, course size, course weekday, students' gender), while others had an impact on how students rate a few of the questions.

The fact that the teacher had access to the results of mid-term evaluations had a positive effect on course evaluation scores for most questions (all except A.1.6 and A.1.7), which is consistent with previous findings (Clemmensen et al., 2013). Course prerequisites and course workload cannot be easily changed during the semester, while such issues as teacher communication and teacher feedback to students can be improved during the semester. Students from the courses where teachers made adjustments to the courses based on the mid-term evaluation results were rewarded with higher SET scores.

The data we analysed are pooled data from all students that evaluated the selected courses in either mid-term evaluations or end-of-term evaluations. The evaluation time point itself only had a significant effect for three questions A.1.3 (Material), A.1.4 (Feedback) and A.1.6 (Workload). At the final evaluations these course aspects were more likely to get lower scores at the end-of-term evaluation than at the mid-term evaluation.

12.7 Conclusions

This study analysed the course-specific and student-specific characteristics that influence students' participation in mid-term and end-of-term student evaluations of teaching at the Technical University of Denmark. An extra mid-term evaluation, identical to the end-of-term, was set up for 35 selected courses in the 6th week of the fall semester in 2010. The end-of-term evaluations were conducted as usual at the 13th week.

On this campus, there was evidence of SET non-response bias both in the mid-term and end-of-term evaluations. Female students, with high GPA, taking the course for the first time were more likely to participate in the course evaluation survey at both time points. However, there were some differences between respondents to the mid-term SET and end-of-term SET. Course-specific characteristics had a different direction or even sign of effect for the two time points.

In addition to non-participation in SET, the effects of the course-specific and student-specific characteristics on the SET scores were investigated. The main conclusion is that even though the overall high achievers (based on GPA), were more likely to participate in the evaluation survey, the GPA itself had little effect on the SET scores. However, the grade obtained on the course was strongly positively correlated with both SET participation and SET scores. The gender bias was found to be present in both mid-term and end-of- term SET participation and in SET ratings.

Bibliography

Robert D. Abbott, Donald H. Wulff, Jody D. Nyquist, Vickie A. Ropp, and Carla W. Hess. Satisfaction with processes of collecting student opinions about instruction: The student perspective. *Journal of Educational Psychology. Special Section: Instruction in Higher Education.*, 82(2):201–206, June 1990.

Philip C. Abrami. *Improving judgments about teaching effectiveness using teacher rating forms.*, volume [Special issue]. New Directions for Institutional Research. 2001.

Philip C. Abrami and Sylvia d'Apollonia. Multidimensional students' evaluations of teaching effectiveness - generalizability of "n = 1" research: Comment on marsh (1991). *Journal of Educational Psychology*, (30):221–227, 1991.

Philip C. Abrami, Sylvia d'Apollonia, and Steven Rosenfield. The dimensionality of student ratings of instruction: what we know and what we do not. *Perry, R.P., Smart J.C., editors: Effective teaching in higher education: research and practice. New York: Agathon Press*, 1997.

Philip C. Abrami, Sylvia dApollonia, and Steven Rosenfield. The dimensionality of student ratings of instruction: what we know and what we do not. *Perry, R.P., Smart J.C., editors: effective teaching in higher education: research and practice. New York: Agathon Press*, pages 385–456, 2007.

Meredith J D Adams and Paul D Umbach. Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. *Research in Higher Education*, 53(5):576–591, August 2012.

Meredith Jane Dean Adams. No evaluation left behind: Nonresponse in online course evaluations. *Doctoral dissertation. North Carolina State University*, 2010.

Shotaro Akaho. A kernel method for canonical correlation analysis. *In Proceedings of the International Meeting of the Psychometric Society (IMPS2001)*, 2006.

Lawrence M. Aleamoni. Typical faculty concerns about student evaluation of teaching. *Techniques for evaluation and improving instruction: New Directions for Teaching and Learning*, (31):25–31, 1987.

Lawrence M. Aleamoni. Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, 13(2):153–166, 1999.

F. N. A. Alhija and B. Fresko. Student evaluation of instruction: What can be learned from students' written comments? *Studies in Educational Evaluation*, 35(1):37–44, 2009.

Vibeke Normann Andersen, Peter Dahler-Larsen, and Carsten Strømbæk Pedersen. Quality assurance and evaluation in denmark. *Journal of Education Policy*, 24(2):135–147, 2009.

Heidi M. Anderson, Jeff Cain, and Eleanora Bird. Online student course evaluations: Review of literature and a pilot study. *American Journal of Pharmaceutical Education*, 69(1), 2005.

Arcanic. A/S.

Raoul Albert 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.) Bolton, M. A.: Anker., 2000.

Lori Price Aultman. An unexpected benefit of formative student evaluations. *College Teaching*, 54(3):251–285, 2006.

Rosemary J. Avery, W. Keith Bryant, Alan Mathios, Hyojin Kang, and Duncan Bell. Electronic course evaluations: Does an online delivery system influence student evaluations? *Journal of Economic Education*, 37(1):21–37, 2006.

Tamara Baldwin and Nancy Blattner. Guarding against potential bias in student evaluations: What every faculty member needs to know. *College Teaching*, 51(1):27–32, 2003.

C. Ballantyne. Why survey online? a practical look at issues in the use of the internet for surveys in higher education. *The Annual Conference of the American Evaluation Association*, November 2000.

Tanya Beran, Claudio Violato, Don Kline, and Jim Frideres. The utility of student ratings of instruction for students, faculty, and administrators: A "consequential validity" study. *Canadian Journal of Higher Education*, 32 (5):49 – 70, 2005.

Tanya Beran, Claudio Violato, and Don Kline. What's the 'use' of student ratings of instruction for administrators? one university's experience. *Canadian Journal of Higher Education*, 17(1):27–43, 2007.

Tanya N. Beran and Jennifer L. Rokosh. Instructors' perspectives on the utility of student ratings of instruction. *Instructional Science*, 37(2):171–184, 2009.

Ronald A. Berk. Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education*, 17(1): 48–62, 2005.

J. B. Biggs and C Tang. Teaching for quality learning at university (3rd ed.). *Maidenhead, Berkshire: McGraw-Hill Education*, 2007.

John Biggs. *Teaching for Quality Learning at University.* Maidenhead, Berkshire: Open University Press., 2003.

Roger B. Bradford. An empirical study of required dimensionality for large-scale latent semantic indexing applications. *17th ACM Conference on Information and Knowledge Management*, page 153–162, 2008.

John C. Braskamp, Larry A.and Ory and David M. Pieper. Student written comments: Dimensions of instructional quality. *Journal of Educational Psychology*, 73(1):65–70, 1981.

Larry A. Braskamp and John C. Ory. Assessing faculty work: Enhancing individual and institutional performance. 1994.

Chris Buckley and Gerard Salton. Stop word list. URL `http://www.lextek.com/manuals/onix/stopwords2.html`.

Charles A. Burdsal and Paul D. Harrison. Further evidence supporting the validity of both a multidimensional profile and an overall evaluation of teaching effectiveness. *Assessment & Evaluation in Higher Education*, 33(5):567–576, 2008.

William E Cashin. Student ratings of teaching: A summary of the research. idea paper no. 20. 1988.

William E. Cashin. Student ratings of teaching: The research revised. *Center for Faculty Education and Development*, September 1994.

William E Cashin. Student ratings of teaching: The research revisited. idea paper no. 32. 1995.

William E. Cashin and Ronald G. Downey. Using global student rating items for summative evaluation. *Journal of Educational Psychology*, 84:563–572, 1992.

John A. Centra. *Determining Faculty Effectiveness. Assessing Teaching, Research, and Service for Personnel Decisions and Improvement.* Jossey-Bass Publications, 1979.

John A. Centra. *Reflective Faculty Evaluations: Enhancing teaching and determining faculty effectiveness.* San Francisco: Jossey-Bass Publications, 1993.

John A. Centra. Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education*, 44(5): 495–518, 2003.

John A. Centra. Differences in responses to the student instructional report: Is it bias? 2009.

John A Centra and F Reid Creech. The relationship between student, teacher, and course characteristics and student ratings of teacher effectiveness. *Project report*, 76(1), 1976.

W. Cerbin and P. Hutchings. The teaching portfolio. *Paper presented at the Bush Summer Institute, Minneapolis, MN.*, June 1993.

Thomas I Chacko. Student ratings of instruction: A function of grading standards. *Educational Research Quarterly*, 8(2):19–25, 1983.

Xi Chen, Han Liu, and Jaime G. Carbonell. Structured sparse canonical correlation analysis. *Proceedings of the 15th international conference on artificial intelligence and statistics*, pages 199–207, 2012.

Yining Chen and Leon B. Hoshower. Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment & Evaluation in Higher Education*, 28(1):71–88, 2003.

Dennis E. Clayson. Student evaluations of teaching: Are they related to what students learn?: A meta-analysis and review of the literature. *Journal of Marketing Education*, 31(1):16–30, 2009.

Line H. Clemmensen, Tamara Sliusarenko, Birgitte Lund Christiansen, and Bjarne K. Ersbøll. Effects of mid-term student evaluations of teaching as measured by end-of-term evaluations. *CSEDU 2013 (5th International Conference on Computer Supported Education)*, 2013.

J. Cohen, P. Cohen, S.G. West, and L.S. Aiken. *Applied multiple regression/correlation analysis for the behavioural sciences.* Mahwah(NJ): Lawrence Erlbaum, 3rd edition, 2003.

Peter A. Cohen. Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education*, 13 (4):321–341, 1980.

Peter A. Cohen. Student rating of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, 51 (3):281–309, 1981.

Jeffrey Coleman and W. J. McKeachie. Effects of instructor/course evaluations on student course selection. *Journal of Educational Psychology*, 72(2):224–226, April 1981.

Alison Cook-Sather. From traditional accountability to shared responsibility: the benefits and challenges of student consultants gathering midcourse feedback in college classrooms. *Assessment & Evaluation in Higher Education*, 34(2):231–241, 2009.

Frank Costin. Do student ratings of college teachers predict student achievement? *Teaching of Psychology*, 5(2):86–88, 1978.

Tena B. Crews and Dylan F. Curtis. Online course evaluations: Faculty perspective and strategies for improved response rates. *Assessment & Evaluation in Higher Education*, 36(7):865–878, 2011.

Fred Damerau. Generating and evaluating domain-oriented multi-word terms from text. *Information Processing and Management*, 29:433–447, 1993.

Sylvia d'Apollonia and Philip C. Abrami. Navigating student ratings of instruction. *American Psychologist*, 52(11):1198–1208, 1997.

Barbara Gross Davis. *Tools for teaching*. John Wiley & Sons, 2009.

Scott Deerwester. Improving information retrieval with latent semantic indexing. *51st ASIS Annual Meeting (ASIS '88)*, 25, October 1988.

Sébastien Déjean and Ignacio González. Package "CCA: Canonical correlation analysis". *CRAN*, 2009.

Curt J. Dommeyer, Paul Baum, and Robert W. Hanna. College students' attitudes toward methods of collecting teaching evaluations: In-class versus on-line. *Journal of Education for Business*, 78(1):11–15, 2002.

Curt J Dommeyer, Paul Baum, Robert W Hanna, and Kenneth S Chapman. Gathering faculty teaching evaluations by in-class and online surveys: their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education*, 29(5):611–623, 2004.

Judy Donovan, Cynthia E. Mader, and John Shinsky. Constructive student feedback: Online vs. traditional course evaluations. *Journal of Interactive Online Learning*, 9(3):283–296, 2006.

Kristina Edström. Doing course evaluation as if learning matters most. *Higher Education Research & Development*, 27(2):95–106, 2008.

C. R. Emery, T. R. Kramer, and R.G. Tian. Return to academic standards: a critique of student evaluations of teaching effectiveness. *Quality Assurance in Education*, 11(1):37–46, 2003.

Bjarne Kjær Ersbøll. Analysing course evaluations and exam grades and the relationships between them. *Proceedings of the Second International Conference on Computer Supported Education, Valencia, Spain, April 7-10*, 1, 2010.

Kenneth A. Feldman. Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher education*, 9(3), 1978.

Kenneth A Feldman. The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education*, 10(2): 149–172, 1979.

Kenneth A. Feldman. Effective collage teaching from the students' and faculty's view: Matched or mismatched priorities? *Research in Higher education*, 28 (4), 1988.

Kenneth A. Feldman. The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher education*, 30(6), 1989a.

Kenneth A Feldman. Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education*, 30(2):137–194, 1989b.

Kenneth A Feldman. College students' views of male and female college teachers: Part ii—evidence from students' evaluations of their classroom teachers. *Research in Higher Education*, 34(2):151–211, 1993.

Kenneth A. Feldman. Identifying exemplary teachers and teaching: Evidence from student ratings. *In R. P. Perry & J. C. Smart (Eds.), Effective teaching in higher education: Research and practice. New York: Agathon Press.*, pages 368–395, 1997.

Kenneth A Feldman. Identifying exemplary teachers and teaching: Evidence from student ratings. In *The scholarship of teaching and learning in higher education: An evidence-based perspective*, pages 93–143. Springer, 2007.

James Felton, John Mitchell, and Michael Stinson. Web-based student evaluations of professors: the relations between perceived quality, easiness and sexiness. *Assessment & Evaluation in Higher Education*, 29(1):91–108, 2004.

James Felton, Peter T. Koper, John Mitchell, and Michael Stinson. Attractiveness, easiness and other issues: student evaluations of professors on ratemyprofessors.com. *Assessment & Evaluation in Higher Education*, 33(1):45–61, 2008.

C.G. Fidelman. *Course Evaluation Surveys: In-class Paper Surveys Versus Voluntary Online Surveys*. Boston College, 2007.

David S. Fike, Denise J. Doyle, and Robert J. Connelly. Online vs. paper evaluations of faculty: When less is just as good. *Journal of Effective Teaching*, 10(2):42–54, 2010.

Ron Fisher and Dale Miller. Responding to student expectations: a partnership approach to course evaluation. *Assessment & Evaluation in Higher Education*, 33(2):191–202, 2008.

Bill & Melinda Gates Foundation. Asking students about teaching. student perception surveys and their implementation, 2012. URL `http://www.metproject.org/`.

Peter W Frey. Validity of student instructional ratings: Does timing matter? *The Journal of Higher Education*, pages 327–336, 1976.

Franklin D. Gaillard, Sonja P. Mitchell, and Vahwere Kavota. Students, faculty, and administrators' perception of students' evaluations of faculty in higher education business schools. *Journal of College Teaching & Learning*, 3(8): 77–90, 2006.

Gerald M. Gillmore, Michael T. Kane, and Richard W. Naccarato. The generalizability of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement*, 15:1–13, 1978.

Abhishek Golugula, George Lee, Stephen R. Master, John E Feldman, Michael D. an Tomaszewski, David W Speicher, and Anant Madabhushi. Supervised regularized canonical correlation analysis: integrating histologic and proteomic measurements for predicting biochemical recurrence following prostate surgery. *BMC Bioinformatics*, 12:483, 2011.

Ignacio González, Sébastien Déjean, Pascal G. P. Martin, and Alain Baccini. Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis. *Journal of Biological Systems*, 17(2):173–199, 2009.

Pamela Gravestock and Emily Gregor-Greenleaf. Student course evaluations: Research, models and trends. *The Higher Education Quality Council of Ontario*, 2008.

William H. Greene. *Econometric Analysis.* Prentice Hall, 5th edition, 2006.

Anthony G. Greenwald. Validity concerns and usefulness of student ratings of instruction. *American Psychologistn*, 52(11):1182–1186, 1997.

Anthony G. Greenwald and Gerald M. Gillmore. Grading leniency is a removable contaminant of student ratings. *Assessment & Evaluation in Higher Education*, 52(11):1209–1217, 1997.

Robert M. Groves and Emilia Peytcheva. The impact of nonresponse rates on nonresponse bias a meta-analysis. *Public Opinion Quarterly*, 72(2):167–189, 2008.

J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, and R. L Tatham. *Multivariate Data Analysis.* Prentice Hall, 6th edition, 2006.

David R. Hardoon and John Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83(3):331–353, 2011.

Nedra Hardy. Online ratings: Fact and fiction. *New Directions for Teaching and Learning*, 2003(96):31–38, 2003.

Suriyati Harun, Suguna K. Dazz, Nurashikin Saaludin, and Wan Suriyani Che Wan Ahmad. Technical lecturers' perception on student evaluation. In *Proceedings of the 7th International Conference on Ubiquitous Information Management and Communication*, ICUIMC '13, pages 55:1–55:8, 2013.

Robert Haskell. Administrative use of student evaluation of faculty. *Education Policy Analysis Archives*, 5, 1997.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2001.

Catherine L. Hatfield and Elizabeth A. Coyle. Factors that influence student completion of course and faculty evaluations. *American Journal of Pharmaceutical Education*, 77(2), 2013.

Suzanne M. Hobson and Donna M. Talbot. Understanding student evaluations: What all faculty should know. *College Teaching*, 49(1):26–31, 2001.

Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artif. Intell. Rev.*, 22(2):85–126, October 2004.

Linda C. Hodges and Katherine Stanton. Changing practices in evaluating reaching. a practical guide to improved faculty performance for promotion/tenure decisions. *Innovative Higher Education*, 31(5):279–286, March 2007.

A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 1933.

Harold Hotelling. The most predictable criterion. *Journal of Educational Psychology*, 26:139–142, 1935.

Harold Hotelling. Relation between two sets of variates. *Biometrika*, 28(3-4): 321–377, 1936.

Paul Isely and Harinder Singh. Do higher grades lead to favorable student evaluations? *The Journal of Economic Education*, 36(1):29–42, 2005.

Tor Aase Johannessen, Kjell Grønhaug, Nils G. Risholm, and Øyvind Mikalsen. What is important to students? exploring dimensions in their evaluations of teachers. *Scandinavian Journal of Educational Research*, 41(2):165–177, 1997.

Richard A. Johnson, Irwin Miller, and John Freund. Miller and freund's probability and statistics for engineers. *Pearson Education, 8th Ed*, 2011.

Trav D. Johnson. Online student ratings: Will students respond? *New Directions for Teaching and Learning*, 2003(96):49–59, 2003a.

Valen E. Johnson. Grade inflation: A crisis in college education. *Springer-Verlag*, 2003b.

C. R. Jones. Nonresponse bias in online sets. *Doctoral dissertation, James Madison University.*, 2009.

Donald W. Jordan. Re-thinking student written comments in course evaluations: Text mining unstructured data for program and institutional assessment. *Doctoral dissertation, California State University.*, May 2011.

Michael T. Kane, Gerald M. Gillmore, and Terence J. Crooks. Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement*, 13(3):171–183, 1976.

Jennifer B. Kasiar, Sara L. Schroeder, and Sheldon G. Holstad. Comparison of traditional and web-based course evaluation processes in a required, team-taught pharmacotherapy course. *American Journal of Pharmaceutical Education*, 66:268–270, 2002.

David Kember, Doris Y. P. Leung, and K. P. Kwan. Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment & Evaluation in Higher Education*, 27(5):411–425, 2002.

Carolin S. Keutzer. Midterm evaluation of teaching provides helpful feedback to instructors. *Teaching of Psychology*, 20(4):238–240, 1993.

Samer Kherfi. Whose opinion is it anyway? determinants of participation in student evaluation of teaching. *Journal of Economic Education*, 42(1):19–30, 2011.

William H. Kilpatrick. The project method. *Teachers College Record*, 19(4): 319–335, 1918.

Christopher Knapper and Patricia Cranton. *Fresh Approaches to the Evaluation of Teaching: New Directions for Teaching and Learning, Number 88*. San Francisco: Jossey-Bass., 2001.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. *MT Summit*, pages 79–86, 2005.

James A. Kulik. Student ratings: Validity, utility, and controversy. *New Directions for Institutional Research*, 2001(109):9–25, 2001.

James A. Kulik and Wilbert J. McKeachie. The evaluation of teachers in higher education. *Review of Research in Education. In F. N. Kerlinger (ed.)*, 3, 1975.

M. C LaForge. Student mood and teaching evaluation ratings. *Journal of The Academy of Business Education*, 4, 2003.

Amy Langville. The linear algebra behind search engines. *The Mathematical Association of America*, December 2005.

Benjamin H Layne, Joseph R Decristoforo, and Dixie Mcginty. Electronic versus traditional student ratings of instruction. *Research in Higher Education*, 40 (2), 1999.

Kim-Anh Le Cao, Pascal G. P. Martin, R. Christele, and P. Besse. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, 10:Article 34, 2009.

LearningLab. DTU. URL `http://www.learninglab.dtu.dk/english`.

Layne Leslie. Defining effective teaching. *Journal on Excellence in College Teaching*, 23, March 2012.

S.E. Leurgans, R. Moyeed, and B.W. Silverman. Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society B*, 55(3): 725–740, 1993.

K. G. Lewis. Making sense of written student comments. *New Directions for Teaching and Learning*, 87:25–32, 2001.

R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55, 1932.

Julie Beth Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1 and 2), 1968.

Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, April 1958.

J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press.*, 1:281–297, 1967.

Christopher D Manning and Hinrich Schutze. Foundations of statistical natural language processing. *MIT Press*, 1999.

Herbert W. Marsh. Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology*, 74:264–279, 1982a.

Herbert W. Marsh. Seeq: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, 52(1):77–95, 1982b.

Herbert W. Marsh. Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, 76:707–754, 1984.

Herbert W. Marsh. Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, 11(3):253 – 388, 1987.

Herbert W. Marsh. Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. *R.P. Perry and J.C. Smart (eds.), The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, pages 319 –383, 2007.

Herbert W. Marsh and Terri L. Cooper. Prior subject interest, students' evaluations, and instructional effectiveness. *Multivariate Behavioral Research*, 16 (1):83–104, 1981.

Herbert W. Marsh and M. Dunkin. Students' evaluations of university teaching: A multidimensional perspective. *Higher education: Handbook on theory and research*, 8:143–234, 1992. Reprinted in R. P. Perry & J. C. Smart (eds.), Effective Teaching in Higher education: Research and Practice (Agathon, 1997) pp. 241-320.

Herbert W. Marsh and Dennis Hocevar. Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education*, 7(4):303 – 314, 1991.

Herbert W. Marsh and Lawrence Roche. The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal*, 30(1):217–251, 1993.

Herbert W. Marsh and Lawrence A. Roche. Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, 52(11):1187, 1997.

Herbert W Marsh and Lawrence A Roche. Effects of grading leniency and low workload on students' evaluations of teaching: popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology*, 92(1):202, 2000.

Herbert W. Marsh and John E. Ware. Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: new interpretations of the dr. fox effect. *Journal of Educational Psychology*, 74:126–134, 1992.

James R. Martin. Evaluating faculty based on student opinions: Problems, implications and recommendations from deming's theory of management perspective. *Issues in Accounting Education*, pages 1079–1094, Nowember 1998.

J. McGourty, K. Scoles, and S. Thorpe. Web-based student evaluation of instruction: Promises and pitfalls. *Paper presented at the 42nd Annual Forum of the Association for Institutional Research: Toronto, ON, Canada*, June 2002.

Wilbert J. McKeachie. Student ratings of faculty. *American Association of University Professors Bulletin*, 55(2):439–444, December 1969.

Wilbert J. McKeachie. Student ratings of faculty: A reprise. *Academe*, 65(6): 384–397, October 1979.

Wilbert J. McKeachie. Student ratings: Their validity of use. *American Psychologist*, 52:1218–1225, 1997.

D. C. Munz and H. E. Munz. Student mood and teaching evaluations. *Journal of Social Behavior and Personalityl*, 12(1):233–242, 1997.

Harry G Murray. Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology*, 75 (1):138, 1983.

D. H. Naftulin, J. E. Ware, and F. A. . Donnelly. The doctor fox lecture: A paradigm of educational seduction. *Journal of Medical Education*, 48:630–635, 1973.

Fadia Nasser and Barbara Fresko. Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education*, 27(2):187–198, 2002.

John Norris and Cynthia Conn. Investigating strategies for increasing student response rates to online-delivered course evaluations. *Quarterly Review of Distance Education*, 6(1):13–29, 2005.

Clifford Nowell. The impact of relative grade expectations on student evaluation of teaching. *International Review of Economics Education*, 6(2):42–56, 2007.

Richard L. Oliver and Elise Pookie Sautter. Using course management systems to enhance the value of student evaluations of teaching. *Journal of Education for Business*, 80(4):231–234, 2005.

John Ory. Faculty thoughts and concerns about student ratings. in k.g. lewis (ed.), techniques and strategies for interpreting student evaluations [special issue]. *New directions for teaching and learning*, 87:3–15, 2001.

John Ory, Larry A. Braskamp, and David M Pieper. Congruency of student evaluative information collected by three methods. *Journal of Educational Psychology*, April 1980.

J. U. Overall and Herbert W. Marsh. Midterm feedback from students: Its relationship to instructional improvement and students' cognitive and affective outcomes. *Journal of Educational Psychology*, 71(6), December 1979.

J. U. Overall and Herbert W. Marsh. Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology*, 72: 321–325, 1980.

David D. Palmer. *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010.

Elena Parkhomenko, David Tritchler, and Joseph Beyene. Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proceedings*, 1 (Suppl 1), 2007.

Mari-Sanna Paukkeri and Timo Honkela. Likey: Unsupervised Language-Independent Keyphrase Extraction. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*, pages 162–165, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

Mari Sanna Paukkeri, Ilari T. Nieminen, Matti Pöllö, and Timo Honkela. A language-independent approach to keyphrase extraction and evaluation. In *In Proceedings of COLING*, 2008.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazineh*, 2(11):559–572, 1901.

Angela R. Penny and Robert Coe. Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of Educational Research*, 74(2): 215–253, 2004.

Martin Porter. Snowball. URL `http://snowball.tartarus.org/`.

Martin Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980.

Stephen R. Porter and Paul D. Umbach. Student survey response rates across institutions: Why do they vary? *Research in Higher Education*, 47(2):229–247, 2006.

Stephen R. Porter and Michael E. Whitcomb. Non-response in student surveys: The role of demographics,engagement, and personality. *Research in Higher Education*, 46(2):127–152, 2005.

James S Pouder. Transformational classroom leadership: a novel approach to evaluating classroom performance. *Assessment and Evaluation in Higher Education*, 33(3):233–243, 2008.

Antti-Tuomas Pulkka and Markku Niemivirta. Predictive relationships between adult students' achievement goal orientations, course evaluations, and performance. *International Journal of Educational Research*, 2012.

William J. Read, Dasaratha V. Rama, and K. Raghunandan. The relationship between student evaluations of teaching and faculty evaluations. *Journal of Education for Business*, pages 189–192, March/April 2001.

John T. E. Richardson. Instruments for obtaining student feedback: a review of the literature. *Assessment & Evaluation in Higher Education*, 30(4):387–415, 2005.

Donald B. Rubin. Multiple imputation for nonresponse in surveys. *New York: Wiley & Sons*, 1987.

James J. Ryan, James A. Anderson, and Allen B. Birchler. Student evaluation: The faculty responds. *Research in Higher Education*, 12(4):317–333, 1980.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523, 1988.

Daniel A. Seiver. Evaluations and grades: A simultaneous framework. *The Journal of Economic Education*, 14(3):32–38, Summer 1983.

P. Seldin. Changing practices in evaluating reaching. a practical guide to improved faculty performance for promotion/tenure decisions. *Bolton, MA: Anker*, 1999.

Eugene P. Sheehan and Tara DuPrey. Student evaluations of university teaching. *Journal of Instructional Psychology*, 26(3), September 1999.

Kari Smith and Miriam Welicker-Pollak. What can they say about my teaching? teacher educators' attitudes to standardised student evaluation of teaching. *European Journal of Teacher Education*, 31(2):203–214, 2008.

Philip L. Smith. The generalizability of student ratings of courses: Asking the right questions. *Journal of Educational Measurement*, 16(2):77–87, 1979.

Arthur Tenenhaus and Michel Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257–284, 2011.

B. Thompson. Canonical correlation analysis: uses and interpretation. *Quantitative applications and social sciences*, 47 of Sage university papers, 1984.

Stephen W. Thorpe. Online student evaluation of instruction: An investigation of non-response bias. *42nd Annual Forum of the Association for Institutional Research*, 2002.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

Douglas F. Vincent. The orgin and development of factor analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 2(2):pp. 107–117, 1953.

Hrishikesh D. Vinod. Canonical ridge and econometrics of joint production. *Journal of Econometrics*, 4(2):147–166, 1976.

Sandra Waaijenborg and Aeilko Zwinderman. Penalized canonical correlation analysis to quantify the association between gene expression and dna markers. *BMC Proceedings*, 1(Suppl 1):S122, 2007.

Howard K. Wachtel. Student evaluation of college teaching effectiveness: a brief review. *Assessment & Evaluation in Higher Education*, 23(2):191–212, 1998.

Bruce A. Weinberg, Belton M. Fleisher, and Masanori Hashimoto. Evaluating methods for evaluating instruction: The case of higher education. *NBER Working Paper No. 12844*, Januarry 2007.

Joakim Westerlund. Class size and student evaluations in sweden. *Education Economics*, 16(1):19–28, 2008.

Wendy Bryce Wilhelm. The relative influence of published teaching evaluations and other instructor attributes on course choice. *Journal of Marketing Education*, 26(1):17–30, 2004.

Daniela M. Witten, Robert Tibshirani, and Sam Gross. Package: Penalized multivariate analysis. *CRAN*, 2009a.

Daniela M. Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostat*, 10(3):515–534, 2009b.

R.E. Wright. Student evaluations of faculty: Concerns raised in the literature, and possible solutions. *College Student Journal*, 40(2):417–422, 2006.

Yuankun Yao and Marilyn L. Grady. How do faculty make formative use of student evaluation feedback?: A multiple case study. *Journal of Personnel Evaluation in Education*, 18:107–126, 2005.

Dimitrios Zeimpekis and Efstratios Gallopoulos. Tmg: A matlab toolbox for generating term-document matrices from text collections. *Grouping multidimensional data, Cambridge, MA: MIT Press*, page 187–210, 2005.

Donald W. Zimmerman. Teacher's corner: A note on interpretation of the paired-samples t test. *Journal of Educational and Behavioral Statistics*, 22 (3):349–360, 1997.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

Hui Zou, Trevor Hastie, and Robret Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15, 2006.