Improving the quality of teaching by utilising written student feedback: A streamlined process

Hujala Maija, Knutas Antti, Hynninen Timo, Arminen Heli

DOI: 10.1016/j.compedu.2020.103965

Please cite the publication as follows:

Hujala, M., Knutas, A., Hynninen, T., & Arminen, H. (2020). Improving the quality of teaching by utilising written student feedback: A streamlined process. Computers & Education, 103965, DOI: 10.1016/j.compedu.2020.103965

IMPROVING THE QUALITY OF TEACHING BY UTILISING WRITTEN STUDENT FEEDBACK: A STREAMLINED PROCESS

ABSTRACT

Currently student feedback is mainly evaluated with quantitative methods since qualitative analysis has been highly effort intensive. In this article, we present a process for tapping into the resource of responses to open-ended feedback questions by using a topic-modelling approach that goes beyond listing modelling outcomes. The objective of this study is to present a streamlined, yet rigorous, process for analysing large amounts of written feedback that connects qualitative findings to existing literature, theories and quantitative feedback. The topic models are created using the Latent Dirichlet Allocation (LDA) method, after which qualitative and quantitative evaluation methods are used to validate the topic outcomes. The proposed process can help educators analyse teaching quality on programme- or institution-wide level, or on single courses with a very large number of students. The process systematizes and combines existing processes, is repeatable, and can serve as a basis for richer analysis for educators. In student evaluation of teaching (SET) research, it advances the state of the art in applied topic modelling by demonstrating how to validate the topics via thematic analysis and by connecting them to theoretical frameworks and quantitative data. Previous topic modelling studies in this field follow mainly descriptive approaches. We demonstrate the process with feedback data collected from 6,087 student evaluations of university courses and confirm that quantitative feedback variables can be used to validate qualitative feedback topic-modelling outcomes and thematic analysis provides a more in-depth explanation of the topics. We additionally find that the proposed topic modelling approach discovers new constructs of SET that cannot be distinguished from quantitative SET measures. The main limitation of the study is that the proposed process is novel and requires further evaluation to establish its full validity.

Keywords: Evaluation methodologies; Data science applications in education; Topic modelling; Student evaluation of teaching; Multilevel analysis

1 INTRODUCTION

Since the 1920s, student evaluation of teaching (SET) has been used to measure teaching quality (Marsh, 1987; Uttl et al., 2017; Wallace et al., 2019), and today SET is the most common method to evaluate faculty's teaching performance in higher education institutions (Clayson, 2009; Hoel & Dahl, 2019; Kember et al., 2002; Spooren, 2010; Spooren & Van Loon, 2012; Wallace et al., 2019). In addition, SETs are used frequently for administrative purposes, such as tenure, promotion and merit-pay decisions on teaching personnel (Spooren et al., 2013).

Student feedback often is collected using both Likert-type items or scales, and open-ended questions. While many statistical methods exist for analysing numerical feedback—such as averages, correlations or linear regression analysis—systematic analysis of responses to open-ended questions is less common. Tapping into the written feedback that students provide is difficult for two reasons. First, masses of open-ended text responses must be collected to distinguish any meaningful themes within the feedback. Second, dealing with text data in masses is, itself, a difficult task. Unlike with numerical feedback, most survey tools do not have tools for automatic processing and evaluation of written responses. Without a systematic process, this can be an overwhelming task, and the potential exists

that the responses to the open-ended questions will be discarded without a proper review, i.e., the feedback that they provide will be wasted.

Automated approaches for processing large numbers of text documents can help with analysing written student feedback. Extant literature has reported mainly positive outcomes. Some of the work can be automated, but as a recent study has noted, *'a degree of human intervention is still required in creating reports that are meaningful and relevant to the context'* (Santhanam et al., 2018, p. 60). In one of the earliest studies in this area, Pan et al. (2009) used text-analysis software and a student feedback sample (sample size N = 556) to develop a method for quantifying students' written comments to increase their usefulness in profiling teachers and teaching. In addition, Abd-Elrahman et al. (2010) used a small student feedback sample (N = 25) to compare the results of human interpretation with automated text analysis. They found that text mining of written comments provides an efficient additional or alternative measure in the course-evaluation process. Also, Stupans et al. (2016) and Shah and Pabel (2019), who identified concepts within written student feedback and compared two student cohorts, emphasised a tool's ability to help identify issues that did not emerge from quantitative SET data. In larger numbers and over several years, Grebennikov and Shah (2013) analysed almost 80,000 comments from study programme feedback surveys from 2001 to 2011 using a text-analytics approach. Their results indicate that by analysing a time series of written feedback via a text-analytics tool, it is possible to discover changes in students' experiences over a given time period.

The latest studies in this field largely deal with proposing algorithms and approaches, both unsupervised and supervised, for detecting topics from students' written comments or classifying them into categories and/or sentiments. For example, Gottipati et al. (2018a) applied both LDA topic models and clustering models to extract topics from the written student feedback and then employed various sentiment mining techniques to classify the students' comments as positive or negative. They concluded that the LDA models can find more relevant topics for the comments compared to the clustering models and, as to the sentiment mining, classification method performs better than the lexicon-based methods. Also, for example, Cunningham-Nelson et al. (2018), Cunningham-Nelson et al. (2019), Pyasi et al. (2018) and Unankard and Nadee (2020) have applied LDA models to detect topics from SET data. As to the supervised methods, McDonald et al. (2019) used a supervised text classification tool (Quantex) to summarise and categorise SET data and compared automated analysis of students' comments with human analysis. They concluded that although the automated methods cannot replace human analysis and evaluation of written comments, they are valuable in assisting in the analysis of students' free text responses. Gottipati et al. (2018b) for their part evaluated rule-based methods and statistical classifiers for extracting explicit suggestions from students' written comments. According to them the statistical classifiers, especially support vector machine and decision tree, provided better classification performance than the rule-based methods. Similar findings were reported by Ibrahim et al. (2019) who compared four machine learning algorithms for filtering assessment related feedback from student's comments. As for sentiment analysis, several studies (Andersson et al., 2018; Baddam et al., 2019; Hew et al., 2020; Onan, 2019; Peng & Xu, 2020; Sengkey et al., 2019) have proposed various approaches for classifying students' written comments as positive, (neutral) or negative. According to, for example, Andersson et al. (2018) the sentiment analysis approach can capture meaningful information from SET, is minimally labour intensive and allows to determine sentiments of a large amount of written responses. Furthermore, Jena (2019) and Nimala and Jebukumar (2019) take one step further and use sentiment mining techniques to model and predict students' emotions, such as joy, trust and anxiety, based on student feedback.

This study contributes to extant literature by developing a streamlined, yet rigorous, process for higher education institutions to analyse masses of responses to open-ended feedback questions. Our aim is not to provide a new method for an individual lecturer to extract topics or sentiments from student feedback given to him/her, but to propose a comprehensive approach for institutions or study programmes to summarise and categorise large amounts of SET data in order to bring new value to the community of teachers and administrative staff. The drawback of the automated text mining compared to manual analysis of students' written comments is loss in the richness of the data. However, we assume that machine learning and traditional qualitative analysis are not alternative methods to analyse large amounts of students' open-ended responses to the feedback questions. Without assistance of the automated text analysis methods the large-scale analysis of written feedback is very easy to dismiss as an extremely time-consuming task.

We propose an unsupervised method (LDA topic modelling) for extracting topics from written student feedback. First, in the absence of training dataset, as it is in our case, an unsupervised method is a natural choice as training a classifier is a laborious task. Second, using a pre-established training set may lead to loss of information and/or ignoring possible new emerging topics as written student feedback is usually very unstructured, informal and situation dependent. In addition, as simply extracting topics or sentiments from student feedback is of little use to educational institutions, we propose naming and describing LDA topics by thematic analysis and compare the results with existing studies in the area. Without linking the results of text-mining to extant literature, the administrative staff and educators cannot have confidence if the topics identified have construct (Campbell & Cook, 1979) and interpretive (Maxwell, 1992) validity, and whether they are relevant to the SET body of knowledge. Furthermore, we apply statistical methods as well to further assess the validity of the topic-modelling outcomes.

Thus, the main research question is: *'How can responses to the open-ended questions from student feedback surveys be analysed to provide validated summaries of student evaluations of teaching?'*

The main research question is divided further into sub-questions, which are listed as follows:

- Which themes emerge from the student feedback via LDA topic modelling?
- How are the themes related to existing knowledge of SET determinants (such as pedagogical issues, course arrangements, student motivation, etc.)?
- How and to what extent are LDA topic-modelling results related to SET dimensions that the close-ended questions measure?

We use data from student feedback surveys (N = 6,087) taken at a Finnish university during the 2016-2017 and 2017-2018 academic years. First, we follow the approach suggested by Finch et al. (2018) and use a topic-modelling method to analyse responses to the open-ended questions. Second, as suggested by Hagen (2018), we thoroughly evaluate and validate topic models' outcomes. The first step in the evaluation process is to use thematic analysis to assign a theme for each topic and compare the themes with existing knowledge on SET determinants. The second step is to use statistical analysis (multilevel regression analysis) to assess the relationships between the outcomes of topic models and the Likert-scale questions from the student feedback instrument used.

The remainder of this paper is organised as follows. Section 2 describes the proposed approach for systematic analysis of written student feedback. The approach is tested in Section 3, the results are discussed in Section 4 and the paper concludes in Section 5.

2 A STREAMLINED APPROACH FOR SYSTEMATIC ANALYSIS OF WRITTEN STUDENT FEEDBACK

In this paper, we present a streamlined approach to systematic analysis of written student feedback based on pre-existing, validated qualitative and quantitative research methods: LDA topic modelling; thematic analysis; and multilevel modelling. What is novel in our approach is combining them into a single analysis pipeline that allows for creating actionable reports from open-ended feedback data. The process combines text mining to pre-sort the open-ended feedback, and a thematic analysis approach rooted in the SET literature to analyse the automatically generated topics. Educators can benefit from this systematic analysis process on a programme or institution level, or in other contexts where open feedback is collected en masse. In this section, we first present an overview of the process (see Figure 1 and Table 1), then detail each step.

The process applies a mixed methods approach, where quantitative and qualitative research techniques, methods, or approaches are combined into a single study (Johnson & Onwuegbuzie, 2004). The specific method we apply in each step is specified in the third column of Table 1. The process can be divided into four main steps: 1) data collection; 2) topic model generation; 3) evaluation of topic-modelling outcomes; and 4) reporting the results. Evaluation of the topic models is divided further into two parallel approaches: qualitative evaluation using thematic analysis and statistical evaluation using multilevel modelling.

[Place Figure 1 approximately here]

Table 1. A brief description of the steps of the proposed approach

Step	**Description**	**Methods Used**	**Expanded in Section**
1. Data collection	Collection of student feedback data	Student feedback surveys, including both Likert-scale and open-ended questions	2.1
2. LDA topic modelling	Categorisation of students' written responses to topics	Pre-processing and cleaning the text data Identifying the optimal number of topics using semantic coherence value Generating topics using the LDA topic-modelling approach	2.2
3. Evaluation of topic models' outcomes	Evaluation of modelled topics' validity	Qualitative and quantitative evaluation methods	2.3
3.1 Qualitative evaluation	Description and naming of the topics Comparison with theory First part of construct validation	Thematic analysis using topics as a starting point Comparison of emergent themes with SET determinants found in extant literature	2.3.1

3.2 Quantitative evaluation	Statistical analysis to examine the relationships between LDA topics and SET dimensions measured with Likert scales. Second part of construct validation.	Multilevel modelling	2.3.2
4. Reporting	Reporting findings as a summary that is understandable without further data analysis and provides a basis for further actions.	-	4

2.1 Data collection

The student feedback data used in this study comes from the feedback surveys carried out at a Finnish university during the 2016-2017 and 2017-2018 academic years. The institution's student feedback questionnaire has been subjected to several revisions during the past few years, giving us a unique opportunity to test LDA topic modelling in the context of student feedback with data collected through two slightly different student feedback questionnaires: one for the academic year 2016-17 and the other for 2017-18. The English versions of the questionnaires are provided in Appendix A.

The first questionnaire (2016-17) comprised seven Likert-scale questions assessing students' motivation, effort put into learning, teacher (two questions), teaching methods, workload and students' perceived learning. In addition, an overall mark for the course, one open-ended question (*'Other feedback about the course (for example, ways to enhance learning during the* course)') and seven background questions were included in the questionnaire.

The second questionnaire (2017-18) comprised five Likert-scale questions assessing students' motivation, effort put into learning, workload, teaching methods and course implementation in relation to perceived learning. Five open-ended questions, one for each Likert-scale item, were included as well: *'What factors affected my level of motivation?'*; *'What factors affected how much I invested in my learning?'*; *'What factors affected the workload?'*; *'My feedback regarding the teaching methods:'*; and *'What factors promoted my learning and how could learning be supported better?'*

The survey questionnaires were sent to students via email after they completed the courses. The surveys mostly were sent to all students enrolled in the courses but considering that some students enrol in courses and do not actually participate, teachers can collect attendance and limit feedback surveys to only those students who attended classes. Responses were collected anonymously and voluntarily from students.

The total number of responses/courses was 9,148/555 in 2016-17 and 8,092/577 in 2017-18. This study is restricted to student feedback written in Finnish, so we included only those responses that contained answers to open-ended question(s) written in Finnish. Thus, the sample size in LDA topic modelling was 6,087, including 2,445 responses collected in 2016-17 and 3,642 responses collected in 2017-18. To address our third research sub-question, we further restricted the sample to courses with three or more student feedback questionnaires filled out. In addition, we excluded PhD courses and courses without teacher(s), such as internships, to increase the sample's homogeneity. The sample sizes in multilevel modelling were 2,323 (2016-17 sample) and 3,496 (2017-18 sample).

2.2 LDA topic modelling

Topic modelling is a statistical text-mining method that can be used to distinguish recurring themes from a set of text documents (Blei, 2012). According to a recent study by Finch et al. (2018), topic modelling is a powerful tool for identifying underlying topics from responses to open-ended survey questions. In addition, for example, Hagen (2018) emphasises topic modelling's ability to identify and extract possible multi-faceted or latent themes.

In this paper, we used the LDA topic-modelling algorithm (Blei et al., 2003) to distinguish topics from student feedback data described in the previous section. LDA is a probabilistic topic-modelling approach based on the Bayesian network model. The LDA algorithm assigns each document (i.e., feedback row) a probability of belonging to a specific topic. Data preprocessing and LDA analysis of the data are described in detail in our recent paper (Author, 2019).

In the pre-processing of the feedback texts we merged the responses to all (five) open-ended questions for the 2017-18 questionnaire. This was because answers could be short, and students often answered only to some questions. Additionally, students do not always differentiate between the different feedback questions, and therefore the question context is not necessarily a reliable indicator, as pointed out by Alhija and Fresko (2009). We also wanted to preserve as much richness as possible in the data for analysis, and therefore decided to treat the open-text answers from different questions as single texts in the LDA analysis.

Below is a short description of the data preprocessing and analysis process used:

1. Downloading student feedback data from an online survey tool into a tabular file format.
2. Sorting the feedback by language and selecting Finnish language responses.
3. Preprocessing and cleaning the data.
    a. Assigning each row of feedback to a document unit.
    b. Converting the documents into a corpus containing unigrams and building a document-term matrix using the R *tm* package (Feinerer et al., 2008).
    c. Cleaning the document-term matrix by
        i. removing non-alphabetic characters
        ii. removing Finnish stopwords
        iii. removing empty documents
        iv. stemming words using the *snowball* algorithm (Bouchet-Valat, 2014) for Finnish language
4. Using the R *stm* library (Roberts et al., 2018) to evaluate the number of topics with semantic coherence quality value (Mimno et al., 2011).
5. Using the modified version of the *NAILS* script (Author, 2015), which utilises the *topicmodels* R package (Grün & Hornik, 2011) to build topic models.
6. Using the *LDAvis* library (Sievert & Shirley, 2015) to visualise models.
7. Exporting analysis outcomes
    a. Summaries of data, including topic probabilities and most characteristic rows per topic
    b. Merging analysis outcomes, including topic probabilities per row of feedback (theta distribution) into the original data table

2.3 Evaluation of topic models' outcomes

While LDA topic modelling can automate many steps in text analysis, it fundamentally is just a probabilistic model in which each topic is defined by a specific set of words that often appear together. There is no guarantee that the generated topics are relevant to the context, and it is the researcher's responsibility to evaluate the topics' validity and assign meaning to content (Brookes & McEnery, 2019; Hagen, 2018).

Hagen (2018) presented a framework in which to train and validate the LDA algorithm using four approaches: 1) rating topic quality and internal coherence, and assigning labels to topics; 2) calculation of computer-human inter-rater reliability; 3) evaluation of external validity; and 4) comparison against manually coded results using the same data. In addition to Hagen (2018), Jacobi et al. (2016) and Brookes and McEnery (2019) have touched on issues related to the validity of LDA topic models' outcomes. Jacobi et al. (2016) demonstrated the use of LDA topic modelling in journalism research, finding that LDA topics mainly corresponded to relevant issues in the discourse in question. However, Brookes and McEnery (2019) had a more critical view of LDA topic modelling's validity. They examined LDA topics' thematic coherence using medical patients' feedback data and analysed the 20 most characteristic feedback types for each topic manually. Their results indicated that many of the topics showed only limited thematic consistency. In addition, Brookes and McEnery (2019) stressed out that validation of the results of topic modelling through qualitative analysis by studying 20 most characteristics texts was laborious and time-consuming.

In this paper, the topics generated by LDA are evaluated using two distinct approaches. First, we conduct a qualitative evaluation through thematic analysis (Braun & Clarke, 2006; Joffe & Yardley, 2004) to define and name the topic clusters. The qualitative evaluation resembles the human coding process presented by Hagen (2018), although it proceeds to a more advanced level by proposing partially automated thematic-analysis process to generate themes based on the topics and relates them to existing SET literature and theories. In addition, the objective of the proposed partially automated thematic analysis process is to address issues concerning laborious qualitative analysis phase raised by Brookes & McEnery (2019). Second, we introduce a novel statistical evaluation approach to further evaluate the descriptive validity of the topic-modelling outcomes and verify the interpretive validity of the qualitative evaluation. According to Finch et al. (2018), an advantage of topic modelling is its ability to create variables that can be used in further data analyses together with other variables. We utilise this ability in the evaluation of LDA topics' validity by connecting topics to the Likert-scale questions from the student feedback instruments used. We accomplish this by statistically examining the relationships between written and numerical student feedback and by comparing the analysis outcomes with the results of thematic analysis. The main question under investigation is to find which same and which different constructs quantitative and partly automated qualitative analyses find from the dataset.

2.3.1 Qualitative evaluation: thematic analysis

The topics generated were subjected to thematic analysis to uncover the underlying descriptors behind the topic categories. Thematic analysis is a 'qualitative research method for identifying, analysing and reporting patterns (themes) within the data' (Braun & Clarke, 2006, p. 79) and has been used widely, including in student feedback analysis (Poulos & Mahony, 2008). It is essentially an iterative, qualitative method for reviewing data that aims

toward increased abstraction. It starts with a row-by-row coding process, and the outcome is a set of themes that describe the phenomenon and their relationships.

Two starting points are available for thematic analysis: inductive and theoretical (Braun & Clarke, 2006). In inductive analysis, codes are generated from the data, and in theoretical analysis, the codes are generated from theories published in extant literature. We followed, with some exceptions, the theory-based thematic analysis framework by Braun & Clarke (2006). A theory-based thematic analysis uses existing theoretical frameworks as a basis for coding. Based on our literature review, we selected the coding families from Grebennikov & Shah (2013) and Pan et al. (2009), as they provide a comprehensive set of teacher- and classroom-related SET codes. As a final step in our thematic analysis process, we compared the discovered themes with the SET determinants presented in previous studies.

Where we diverge from a fully manual thematic analysis approach is that we pre-sorted and clustered the data using topic modelling, reducing the amount of manual work involved in searching for themes. Instead, we used thematic analysis to the most characteristic documents from each topic category assigning themes to each cluster, then connecting the themes to existing theories, as proposed by Finch et al. (2018). Topic modelling has been used to automate parts of thematic analysis in other fields, e.g., by Klein et al. (2015). The partially automated thematic analysis process is summarised in Table 2.

Table 2. Partially automated thematic analysis process

Phase and its description
Phase 1: Getting familiarised with topics generated by LDA topic-modelling process
Phase 2: Generating initial codes using a theory-driven approach
Phase 3: Searching for themes
Phase 4: Reviewing themes
Phase 5: Defining and naming themes
Phase 6: Making comparisons with SET determinants presented in the existing SET literature

Combining topic modelling and thematic analysis into a mixed methods approach allows benefiting from best parts from both quantitative and qualitative research approaches. Topic modelling allows pre-sorting large document sets into topic categories and finding most characteristic documents from each category. Thematic analysis in turn provides a systematic method for researchers to investigate topics and assign themes to them based on expert human knowledge. Furthermore, in Section 3 we evaluate if starting the thematic analysis with the most characteristic documents in each topic allows reaching analysis saturation faster.

2.3.2 Quantitative evaluation: multilevel modelling

After the thematic analysis was completed, we further evaluated the LDA topics' validity statistically by examining the relationships between written and numerical student feedback. We applied multilevel regression analysis (multilevel modelling) to explore the associations between the results from the LDA topics and students' SET ratings measured with the Likert scales from the student feedback instruments. Figure 2 illustrates the data used in these analyses. The multilevel models' dependent and independent variables, as well as the analysis method, are described below in detail.

[Place Figure 2 approximately here]

Dependent variables: topic probabilities

The LDA topic-modelling algorithm produces the probability of belonging to each of the topics for each of the written feedbacks in the sample. In our case, these so-called topic probabilities represent the probability of each row of written feedback belonging to each of the topics generated (see Figure 2). Finch et al. (2018) recommend the use of topic probabilities instead of, for example, the most likely topic of the text when interpreting the results from LDA topic modelling. With the help of topic probabilities, it is possible, for example, to analyse the texts' homogeneity within a topic (Finch et al., 2018).

In this study, we go one step further regarding utilisation of topic probabilities and use them to evaluate LDA topics' validity, as they serve as our multilevel regression models' dependent variables. The idea here is that if the Likert scales of SET used in our student feedback instruments can predict topic probabilities (at least to some extent), the LDA topic-modelling algorithm can distinguish meaningful and SET-related topics. In addition, by predicting topic probabilities with the Likert scales, we can determine whether and to what extent the topics found through the LDA algorithm are consistent with numerical measures of SET, and to what extent they provide unique information not covered in numerical feedback.

Independent variables: SET dimensions measured with Likert scales

The student feedback questionnaires used in data collection included several Likert-scale items measuring various SET aspects. According to preliminary correlation analyses, many items were correlated highly with each other, indicating the possibility of latent SET constructs underlying them. Thus, we used exploratory factor analysis (principal factor analysis followed by promax rotation) to identify possible unobserved factors behind the items. A scree plot of eigenvalues was used to determine the optimal number of factors.

In the 2016-17 sample, the following two factors were identified:

Motivation. Two items had high loadings on this factor: *'My motivation on this course was (1 = Very low; 5 = Very high)'* (0.701) and *'I put effort into my learning on this course (1 = Very little; 5 = Very much)'* (0.711). The items were averaged together, creating a measure of students' motivation with acceptable internal consistency (Cronbach's α = 0.789).

Perceptions of teaching. Three items had high loadings on this factor: *'The teaching skills of the teacher(s) supported my learning during the course'* (0.856), *'The expertise of the teacher(s) supported my learning during the course'* (0.842) and *'The teaching methods used supported my learning during the course'* (0.561). The response scale ranged from 1 = *'I do not agree at all'* to 5 = *'I fully agree'*. The items exhibited good reliability (Cronbach's α = 0.875), and they were averaged together to form a measure of students' perceptions of teaching.

Two items – *'The workload for this course in relation to other courses of equal credit was (1 = Much lighter; 5 = Much heavier)'* and *'The course promoted my learning (1 = Very little; 5 = Very much)'* – did not load on the factors and were included in the multilevel models as single-item measures of students' *perceived workload* and *perceived learning*.

Two factors were identified from the 2017-18 sample as well:

Motivation. Two items had high loadings on this factor: *'My motivation in this course was (1 = Very low; 5 = Very high)'* (0.639) and *'I invested in my learning on this course (1 = Very little; 5 = A great deal)'* (0.654). The items were averaged together and formed a measure of students' motivation with acceptable internal consistency (Cronbach's α = 0.782).

Perceptions of the implementation. Two items had high loadings on this factor: *'The teaching methods used on the course supported my learning (1 = Strongly disagree; 5 = Strongly agree)'* (0.719) and *'The manner in which the course was implemented helped me achieve the learning outcomes of the course (1 = Very poorly; 5 = Very well)'* (0.704). The items exhibited good reliability (Cronbach's α = 0.836), and they were averaged together to form a measure of students' perceptions of the course's implementation.

One item, *'The workload relative to the study credits awarded was (*1 = *'Very light'*; 5 = *'Very heavy)',* did not load on any of the factors and was used as a single-item measure of students' *perceived workload* in multilevel regression analyses.

Analysis

Observations of the student feedback data are not fully independent because students are clustered within courses, and courses are clustered within study programmes (see Figure 3). Ignoring data clustering may lead to underestimated standard errors of regression coefficients and, thus, overly small p-values. We took the clustering into account by applying multilevel modelling instead of the traditional ordinary least squares (OLS) estimation method (Snijders & Bosker, 1999).

[Place Figure 3 approximately here]

Multilevel models allow for residual components at all levels – in our case, at study-programme level (v_k), course level (u_{jk}) and student level (e_{ijk}).

We fitted the random intercept models,

$$y_{ijk} = beta_0+beta_1x_{1ijk}+beta_2x_{2ijk}+...+beta_nx_{nijk}+v_k+u_{jk}+e_{ijk} \quad (1),$$

to the student feedback data to answer the third research sub-question (one model for each of the topics found). Dependent variable *y* is the topic probability of a certain topic calculated during LDA topic modelling. Independent variables x_1-x_n include SET variables presented above. For example, x_{1ijk} is the student's *i* self-reported motivation on course *j* organised by study programme *k.*

We used Stata/SE 15.1 software and maximum-likelihood estimation. Academic years 2016-2017 and 2017-2018 were analysed separately due to differences in student feedback instruments and, thus, independent variables used.

3 RESULTS

In this section, we demonstrate the proposed process presented in Section 2.

3.1 Discovered SET topics

After examining the local maximums in the semantic coherence results, we proceeded with a topic model with six topics. The topic-modelling results that we discovered are summarised in Table 3. A more thorough explanation of the SET LDA topic-modelling process and findings is provided in (Author, 2019). The topics' validity is evaluated further in the next two sections using thematic analysis and multilevel modelling.

Table 3. The eight most characteristic words for each topic.

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**
exam	student	final project	exercise	issues	topic
good	teacher	time	do work	complete	good
exercises	everything	group	always	part	learn
lecture	feedback	done	exercise class	time	learning material
moodle	part	period	week	example	lecturer
material	questions	workload	correct	felt	interesting
better	own	weekly exercises	even	time	suitable
homework	even	others	them	example	motivation

3.2 Qualitative evaluation: thematic analysis of the discovered topics

The word lists presented in Table 3 are of little use to anyone analysing student feedback because not enough information is available, for example, to draw any conclusions or make choices based on them. In fact, as discussed in Section 2.3, there is no guarantee that the generated topics are pertinent to the examination of SET. Thus, to assign meaning to the topics, we used thematic analysis to uncover the underlying descriptors behind the topics. In addition, we compared the discovered themes with previous studies' findings.

The first three authors were involved in the thematic analysis process. We started with **Phase 1** of the process presented in Table 2, getting familiarised with each topic generated by the probabilistic LDA topic-modelling process. We then sorted written feedback into the most likely topics based on their topic probabilities.

Phase 2, generating initial codes, was performed using a theory-driven approach. We selected a set of codes by using the coding families created by Grebennikov & Shah (2013) and Pan et al. (2009). These codes then were applied to the 12 most characteristic feedback texts for each topic, for a total of 72 analysed feedbacks. A summary of codes assigned to the feedbacks for each topic is presented in Table B1 in Appendix B.

In the coding process the authors categorised each of the sample texts according to the selected coding family. Each text was marked to belong in all the suitable categories, and each text would either belong to a category or not. The authors individually conducted the initial coding. As the coding scheme was binary, we did not see fit to calculate inter-rater reliability indicator figures because the coding required very little interpretation (N. McDonald

et al., 2019); However, in order to avoid oversights caused by a single-coder approach, the coding results were cross-checked by the other two authors in the group.

Phase 3, searching for themes, re-focussed the analysis from individual codes to the broader analysis of discovering themes. Essentially, the individual codes in this phase were viewed as groups and how they could be combined into overarching themes. These themes were refined further in **Phase 4**, in which alternative candidate themes and options with most supporting data were selected. Furthermore, in Phase 4, we validated the themes by grounding them into the original data set and verifying that the themes described most feedbacks in each topic. In **Phase 5**, we described the topics and gave each a succinct name. In addition, as a final step in our thematic-analysis process, we linked the discovered themes to the SET determinants presented in the previous literature on SET.

The themes found in Phase 5 are described and related to existing studies in Table 4. As shown, each topic has a coherent theme, and similar constructs can be found in extant literature. The main themes were good content, dissatisfaction with personnel or the course, workload, stress and good teaching methods. When considering similarities and differences in the topics, course content is a topic of discussion in five of six topics, and teaching methods are discussed with half the topics. Only two topics are clearly positive or negative, while other topics bring up negative and positive themes.

Table 4. Comparing topics and themes with SET dimensions found in extant literature

Topic	**Theme**	**Description**	**Similar constructs in extant literature**
Topic 1	Good content; improvements in arrangements	Students are happy with the content and teaching but want improvements in how the course is delivered or paced.	Course content (Alhija & Fresko, 2009; Brockx et al., 2012) Teaching methods (Alhija & Fresko, 2009; Carle, 2009; Lowenthal et al., 2015) Teacher's teaching skills/experience (Alhija & Fresko, 2009; Pan et al., 2009; Stewart, 2015)
Topic 2	Severe, emotional dissatisfaction	Expressions of severe dissatisfaction, especially with course arrangements, assessments and the lecturer's teaching methods or content. Emotional feedback.	Assessments (Brockx et al., 2012) Course content (Alhija & Fresko, 2009; Brockx et al., 2012) Students' expectations (Wachtel, 1998) Teaching methods (Alhija & Fresko, 2009; Carle, 2009; Lowenthal et al., 2015) Teacher's teaching skills/experience (Alhija & Fresko, 2009; Pan et al., 2009; Stewart, 2015)
Topic 3	Dissatisfaction with course	Some positive notes, but dissatisfaction regarding how the course was arranged, especially with clarity, workload and time management.	Course rigor (difficulty/workload/course pace) (Alhija & Fresko, 2009; Centra, 2003; Clayson, 2009; Marsh, 2001; Marsh & Roche, 2000; Remedios & Lieberman, 2008; Ting, 2000) Course scheduling (Alhija & Fresko, 2009; De Witte & Rogge, 2011; Meng Tan & Chye Koh, 1997; Wachtel, 1998)

Topic 4	Dissatisfaction with workload	Feedback about disproportionate workload and instructor or coursework time management. Some feedback about good course content despite the workload.	Course content (Alhija & Fresko, 2009; Brockx et al., 2012) Course rigor (difficulty/workload/course pace) (Alhija & Fresko, 2009; Centra, 2003; Clayson, 2009; Marsh, 2001; Marsh & Roche, 2000; Remedios & Lieberman, 2008; Ting, 2000)
Topic 5	Interesting, but challenging, content; stressful.	Mixed feedback: positive overall feedback from course content and assignment, but negative feedback about the course workload and challenging content.	Course content (Alhija & Fresko, 2009; Brockx et al., 2012) Course rigor (difficulty/workload/course pace) (Alhija & Fresko, 2009; Centra, 2003; Clayson, 2009; Marsh, 2001; Marsh & Roche, 2000; Remedios & Lieberman, 2008; Ting, 2000)
Topic 6	Good teaching methods	Positive feedback, especially regarding interesting content and teaching methods; some minor negative feedback on time management.	Course content (Alhija & Fresko, 2009; Brockx et al., 2012) Teaching methods (Alhija & Fresko, 2009; Carle, 2009; Lowenthal et al., 2015) Course scheduling (Alhija & Fresko, 2009; De Witte & Rogge, 2011; Meng Tan & Chye Koh, 1997; Wachtel, 1998)

In order to evaluate the possible impact of feedback questionnaires on the results of LDA topic modelling, we placed each individual feedback document into one of the topic categories based on the highest topic probability. The percentage shares of feedbacks assigned to topics are presented in Figure 4.

[Place Figure 4 approximately here]

As shown in Figure 4, there is a notable difference in the shares of feedbacks placed into Topics 2 and 6 between the academic years but otherwise the outcomes are almost similar. The academic year 2016-17 has a larger share of feedback related to “severe, emotional dissatisfaction” (Topic 2) whereas in the academic year 2017-18, the larger share of feedback is related to Topic 6, i.e., “good teaching methods”. Thus, the five open-ended questions used in the academic year 2017-18 seem to have led to somewhat more positive feedback compared to the year 2016-17 when only one open-ended feedback question was used. What comes to the content of the topics, when comparing the descriptions of the topics (see Table 4) with the open-ended questions used (see Section 2.1 or Appendix A), it is evident that the topics emerged are not similar to the feedback questions used. The topics are multifaceted and cover a large range of issues not specified in the questions. All in all, it is likely that the feedback questionnaires used have affected the results of the LDA topic modelling. However, it also appears, that the effect is more pronounced when it comes to the emotional value of the feedback compared to the content of the topics.

3.3 Quantitative evaluation: multilevel modelling of topic probabilities

Multilevel modelling's primary objective is to evaluate LDA topics' validity further. Thematic analysis, although done using guidelines, is a subjective method, and the researcher's perspective affects the analysis. Statistical analysis – in this case, multilevel modelling – provides more objective and reproducible results that may be used to verify results from the thematic analysis. Furthermore, multilevel regression analysis is used to examine whether the LDA topics can provide unique information not covered by the Likert-scale questions.

We fitted six three-level random intercept models (1), one for each of the topics, to the feedback data to find out which SET dimensions measured by Likert scales (if any) are associated with the topic probabilities. Descriptive statistics of and correlations between dependent and independent variables are provided in Appendix C, in Tables C1 and C2. The complete estimation results for the random intercept models are provided in Appendix C as well (see Tables C3 and C4). A summary of the estimation results is presented in Table 5 below. Next, we go through the estimation results and compare them with the results from thematic analysis.

Table 5. The associations between written (measured with topic probabilities) and numerical (SET dimensions measured by the Likert scales) student feedback.

2016-17 sample	***Dependent variables (topic probabilities)***					
Independent variables (Likert scales)	Topic 1 (%) Good content; improvements in arrangements	Topic 2 (%) Severe, emotional dissatisfaction	Topic 3 (%) Dissatisfaction with course	Topic 4 (%) Dissatisfaction with workload	Topic 5 (%) Interesting, but challenging content; stressful	Topic 6 (%) Good teaching methods
Motivation	n.s.	+	n.s.	n.s.	n.s.	-
Perceived workload	n.s.	n.s.	n.s.	-	n.s.	n.s.
Perceptions of teaching	n.s.	-	n.s.	n.s.	n.s.	+
Perceived learning	n.s.	-	n.s.	n.s.	n.s.	+
R^2 value	0.016	0.056	0.009	0.018	0.000	0.118
2017-18 sample	***Dependent variables (topic probabilities)***					
Independent variables (Likert scales)	Topic 1 (%) Good content; improvements in arrangements	Topic 2 (%) Severe, emotional dissatisfaction	Topic 3 (%) Dissatisfaction with course	Topic 4 (%) Dissatisfaction with workload	Topic 5 (%) Interesting, but challenging content; stressful	Topic 6 (%) Good teaching methods
Motivation	n.s.	n.s.	-	+	-	+
Perceived workload	-	n.s.	+	n.s.	n.s.	-
Perceptions of implementation	+	-	n.s.	-	-	+
R^2 value	0.030	0.053	0.057	0.049	0.010	0.146

+/- = the estimated coefficient is positive/negative and statistically significant at $p < 0.05$
n.s. = the estimated coefficient is insignificant at the 5% significance level
R^2 is a Level 1 (i.e., student level) explained proportion of variance (Snijders & Bosker, 1999)

2016-17 sample

As shown in Table 5, Likert scales of SET can predict the topic probabilities of topics 2, 4 and 6. When it comes to Topics 2 and 6, the estimation results seem to be in line with the outcomes of the thematic analysis: Students' motivation, students' perceptions of teaching and students' perceived learning contribute significantly to the topic probabilities of Topics 2 and 6. The estimation results indicate that the higher the student's motivation and the more dissatisfied he or she is with teaching and learning, the higher the probability that his or her written feedback is related to Topic 2 (*'Severe, emotional dissatisfaction*'). On the other hand, the lower the student's motivation and the more satisfied he or she is with teaching and learning, the higher the probability that his or her written feedback is related to Topic 6 (*'Good teaching methods')*.

Instead, as to Topic 4, the multilevel modelling's results seem to differ from the results of the thematic analysis. As shown in Table 5, the student's perceived workload is the only statistically significant regressor of the topic probability of Topic 4. The relationship between perceived workload and topic probability is negative, indicating that the higher the student's perceived workload, the lower the probability that his or her written feedback is related to Topic 4. However, given that Topic 4 is named *'Dissatisfaction with workload'*, one would expect that higher values of perceived workload lead to higher, not lower, topic probabilities of Topic 4.

Regarding Topics 1, 3 and 5, the Likert scales are not associated with the topic probabilities, indicating that a student's numerical feedback cannot predict whether his or her written feedback falls into these topic categories. In other words, Topics 1, 3 and 5 seem to provide information about SET not covered by the Likert-scale questions used during the 2016-17 academic year. In addition, as for Topics 2, 4 and 6, the models' R^2 values are 0.056, 0.018 and 0.118, respectively, i.e., the Likert scales can explain only small proportions (1.8%-11.8%) of the variances in topic probabilities. Thus, it seems that Topics 2, 4 and 6 provide lots of information not captured by the Likert-scale questions as well.

2017-18 sample

As shown in Table 5, the Likert scales are associated with all six topic probabilities. Also, it seems that the Likert scales' estimated coefficients are mostly in line with the results of the thematic analysis. For example, with Topic 2, the more dissatisfied the student is with the course's implementation, the higher the probability that his or her written feedback is related to this topic (*'Severe, emotional dissatisfaction*'). On the other hand, the lower the student's motivation and the higher the perceived workload, the higher the probability that the student's written feedback will be related to Topic 3 (*'Dissatisfaction with course'*). Also, as for topics 1, 5 and 6, the estimation results correspond fairly well with the results of the thematic analysis. The lower the student's perceived workload and the more satisfied he or she is with the course's implementation, the higher the probability that his or her written feedback is related to Topic 1 (*'Good content; improvements in arrangements*'). Furthermore, the lower the student's motivation and the more dissatisfied he or she is with the course's implementation, the higher the probability that his or her written feedback is related to Topic 5 ('*Interesting, but challenging content; stressful').* Finally, the estimated coefficients indicate that the higher the student's motivation, the lower the student's perceived workload, and the more satisfied he or she is with the course's implementation,

the higher the probability that his or her written feedback is associated with Topic 6 ('*Good teaching methods*').

However, the results of the multilevel regression analysis regarding Topic 4 seem to contradict the findings of the thematic analysis, at least to some extent, also in this sample. According to Table 5, the higher the student's motivation and the more dissatisfied he or she is with the course's implementation, the higher the probability that his or her written feedback is related to Topic 4 (*'Dissatisfaction with workload'*). However, despite the name of the topic referring to the workload, the student's perceived workload is not associated with Topic 4 in the multilevel regression analysis.

All in all, it seems that the numerical feedback that a student provides is, at least to some extent, able to predict his or her written feedback. However, the R^2 values of the models predicting topic probabilities are relatively low (1.0% - 14.6%), indicating that the Likert scales can explain, at best, only 14.6% (see Topic 6 in Table 5) of the variances in topic probabilities. Thus, as with the 2016-17 sample, it seems that the LDA topics also can provide unique information not covered by the Likert-scale questions.

Summary of the statistical evaluation

In summary, it seems that the Likert scales were able to predict most topics' probabilities, further confirming that the LDA algorithm was able to extract SET-related topics. Furthermore, apart from Topic 4, the results of the multilevel analysis verified the results of the thematic analysis. Thus, it seems that the Topic 4 should be discarded as it appears to be invalid or, at least, further investigated before reporting the findings.

In addition, the results of the statistical evaluation suggest that LDA topics extracted from the textual student feedback can provide lots of unique information not covered by the Likert-scale questions.

4 DISCUSSION

This study's objective was to evaluate the use of topic modelling for analysing written student feedback. To achieve this goal, we established a streamlined process for examining masses of open feedback using a combination of topic modelling, thematic analysis, and multilevel modelling. As a result, we present a reproducible process for handling masses of feedback, which also is rigorously rooted in SET literature.

4.1 Relating the contributions to literature

We extended the state of the art by presenting and demonstrating a mixed methods analysis process and validation methods that go beyond the current use of machine learning methods in SET literature (Gottipati et al., 2017; Nitin et al., 2015; Onan, 2019; Sengkey et al., 2019; Srinivas & Rajendran, 2019) that has concentrated on providing new methods and algorithms to extract topics or sentiments from student feedback. Furthermore, we extend the current topic modelling validation approaches from literature (Brookes & McEnery, 2019; Hagen, 2018; Jacobi et al., 2016) by demonstrating that the combination of thematic analysis and statistical evaluation via regression analysis can be used as an evaluation method. It appears that until now thematic analysis has been used in SET literature to summarise and

categorise written student feedback by, for example, Langan et al. (2017), but has not yet been applied in evaluating topic-modelling outcomes. Partly automated thematic analysis used in this study has the benefit of naming and describing the LDA topics relatively easy and, thus, making them more comparable to existing literature. Validation by linking the discovered topics to the previous studies is essential, as LDA topic modelling is an unsupervised machine learning method, the results of which always need validation (Hagen, 2018).

We present how the outcomes of LDA topic modelling can be further validated by confirmatory statistical methods. We accomplished this by applying multi-level regression analysis and predicting topic probabilities with the Likert scales of student feedback instruments. From a mixed-methods perspective, additional validation was achieved by linking close-ended and open-ended questions to support the construct (Campbell & Cook, 1979), descriptive, and interpretive validity of qualitative analysis (Maxwell, 1992). The statistical evaluation method proposed can demonstrate i) which qualitative findings are supported by close-ended questions, ii) situations where text and Likert scale findings are in conflict and require further investigation, and iii) which open-ended questions find novel constructs. Hew at al. (2020) have linked results provided by machine learning methods with the Likert scales to predict MOOC satisfaction, but, to our knowledge, our study is the first to use this kind of approach for evaluation and validation of LDA topics.

Our third contribution to the state of the art is exploring the practicality of applying qualitative analysis to topic modelling in specific cases of SET, even though it was not found practical in the field of discourse studies (Brookes & McEnery, 2019). We agree with the arguments presented by Brookes & McEnery (2019) that topic word-lists alone are not sufficient to infer themes present in a topic. However, analysing SET at degree program or higher level involves large datasets that would be impractical to analyse by hand. Therefore, qualitative student feedback is often in the danger of being ignored. Combining certain steps of topic modelling and thematic analysis allows pre-sorting topics and algorithm-guided sampling, making thematic analysis practical in larger datasets and speeding up the analysis, addressing some of the concerns of practicality by Brookes & McEnery (2019). In this, our findings agree with the findings by J. McDonald et al. (2019), where they found a supervised learning machine classifier useful for summarizing and categorizing SET. In the study by J. McDonald et al. (2019), they found that around 140 – 175 human-coded rows were required to achieve sufficient accuracy. In our case, we had similar beneficial results with unsupervised LDA algorithm and human labelling, requiring about 70 coded rows. If the dataset is small, we agree that the automated step can be skipped, following for example a thematic analysis approach presented by Langan et al. (2017) or Poulos & Mahony (2008).

4.2 Reflecting on the proposed process and its implications for practise

Our approach is of value to practitioners, as current SET analysis methods emphasise quantitative data. LDA topic modelling and subsequent thematic analysis allow for extracting richer meaning from written feedback without requiring teaching and administrative staff to read every single line of feedback. The approach is systematic and repeatable, and provides findings in quantitative or tabular formats, allowing SET analysis outcome comparisons between units or years. For example, during the analysis process, we extracted theta distribution of topics, or topic probabilities, for each individual document. This allows calculating prevalence of topics on different metadata categories, such as comparing severe emotional dissatisfaction (Topic 2) between courses and/or degree programmes. Comparing such feedback allows, for example, evaluating if freshman years have a good onboarding

process. The proposed analysis process also extracts quantitative indicators from the qualitative data, allowing for their subsequent use in statistical analysis or learning analytics algorithms. For example, the data provided by this process could be used as training material for supervised machine learning algorithms.

Compared to a purely automatic process, we found the evaluation phase to be a valuable step in the process, as topic modelling, by itself, does not provide an explanation of data, only some possible clusters of similar feedbacks. When it comes to qualitative evaluation, the LDA topic modelling helped find initial themes and made it possible to apply a thematic-analysis type of approach to massive data sets. In our demonstration of the process, we found that combining these methods, LDA topic modelling, thematic analysis, and multilevel regression analysis, into a mixed methods approach enabled increased speed and depth of analysis compared to individual approaches. Topic modelling -based text mining enabled fast pre-processing and initial sorting whereas thematic analysis added depth to the topics that pure automatic methods cannot do. The quantitative evaluation using multilevel regression analysis was applied to verify results of the thematic analysis and, thus, to add validity to the data analysis process. In summary, we get the best of both worlds – benefiting from giving human insight to topics and using computational and statistical methods to assist in the most time-consuming parts and in the evaluation of the results.

We found that there are benefits to at least two distinct stakeholders, one for faculty managing the degree programme and other for teachers in charge of courses. For example, the process for SET is centralized in the university providing data for the study – with degree program management and student organizations collecting feedback surveys and providing the feedback to teachers after anonymization. The university has a regular quality assurance process as a part of the external certification process for degree programs. Currently open-ended data plays a small role due to difficulties in processing and summarizing it. Additionally, data from this centralized process can be shared to individual teachers, further saving analysis effort. Adding an analysis step to enhance the data, such as one presented in this paper, would add value both to senior faculty and all teachers.

Creating our process, when developed from start to finish, took approximately four days. Now when the analysis script, step by step instructions, and supporting materials are ready, the entire process could be completed in one or two days. If the dataset were analyzed only, for example with thematic analysis, reviewing the 18000 rows of feedback with qualitative data analysis tools, the process would take much longer time. Since the process is assisted by computational methods and the topic modeling step selects a fixed number of most characteristic rows for qualitative evaluation, the process length is not significantly increased by the amount of data being analyzed. Due to the near-constant analysis time, our process is more beneficial in large datasets, where manual qualitative analysis would be far too laborious.

5 CONCLUSIONS

As the main contribution of this paper, we presented a comprehensive process for analysing masses of responses to the open-ended student feedback questions and a novel method of validating the topic-modelling findings, explaining the topics with thematic analysis, then using statistical analysis for further evaluation of the results. The proposed approach will make it easier to process and analyse large amounts of open-ended responses that often are more informative, but less analytical compared with closed-ended questions. This way the information provided in written student feedback could be used more effectively on a

higher-education institution or study programme level, or by lecturers on very large courses such as MOOCs. Our process adds to the state of the art by continuing lines of research by Hagen (2018), Brookes & McEnery (2019) and Jacobi et al. (2016) on evaluating and validating topic-modelling outcomes, and demonstrating how to apply the proposed evaluation methods in practise.

As a secondary contribution, we provided an example to demonstrate and validate the process. In this demonstration, we followed the approach that is common, for example, in the information system literature for validating processes (Pries-Heje et al., 2008; Venable, 2006). Our demonstration provided an initial, naturalistic evaluation of the process utility and, thus, validation of the process. However, the process that we presented in this paper is novel and requires further evaluation to establish its full validity, which is one of its main limitations. Furthermore, we used written student feedback in only one language in our analysis. In future work, the process should be improved to accommodate multilingual, open-ended feedback. In future research, the process should be applied and evaluated in different cultural and organisational contexts. In addition, the proposed process is not limited to the analysis of student evaluation of teaching but is easily adapted to other surveys that have both open- and closed-ended questions.

REFERENCES

Author (2015). To be added following double-blind review.

Author (2019). To be added following double-blind review.

Abd-Elrahman, A., Andreu, M., & Abbott, T. (2010). *Using text data mining techniques for understanding free-style question answers in course evaluation forms*. *9*(1), 12.

Alhija, F. N.-A., & Fresko, B. (2009). Student evaluation of instruction: What can be learned from students' written comments? *Studies in Educational Evaluation*, *35*(1), 37–44. https://doi.org/10.1016/j.stueduc.2009.01.002

Andersson, E., Dryden, C., & Variawa, C. (2018). Methods of Applying Machine Learning to Student Feedback Through Clustering and Sentiment Analysis. *Proceedings of the Canadian Engineering Education Association (CEEA)*. https://doi.org/10.24908/pceea.v0i0.13059

Baddam, S., Bingi, P., & Shuva, S. (2019). *Student Evaluation of Teaching in Business Education: Discovering Student Sentiments Using Text Mining Techniques*. *13*(3), 14.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Bouchet-Valat, M. (2014). *SnowballC: Snowball Stemmers Based on the C "libstemmer" UTF-8 Library version 0.6.0 from CRAN*. https://rdrr.io/cran/SnowballC/

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brockx, B., Van Roy, K., & Mortelmans, D. (2012). The Student as a Commentator: Students' Comments in Student Evaluations of Teaching. *Procedia - Social and Behavioral Sciences*, *69*, 1122–1133. https://doi.org/10.1016/j.sbspro.2012.12.042

Brookes, G., & McEnery, T. (2019). The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies*, *21*(1), 3–21. https://doi.org/10.1177/1461445618814032

Campbell, D. T., & Cook, T. D. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Rand McNally College Publishing Company Chicago.

Carle, A. C. (2009). Evaluating college students' evaluations of a professor's teaching effectiveness across time and instruction mode (online vs. face-to-face) using a multilevel growth modeling approach. *Computers & Education*, *53*(2), 429–435. https://doi.org/10.1016/j.compedu.2009.03.001

Centra, J. A. (2003). Will Teachers Receive Higher Student Evaluations by Giving Higher Grades and Less Course Work? *Research in Higher Education*, *44*(5), 495–518. https://doi.org/10.1023/A:1025492407752

Clayson, D. E. (2009). 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. https://doi.org/10.1177/0273475308324086

Cunningham-Nelson, S., Baktashmotlagh, M., & Boles, W. (2018). Visually exploring sentiment and keywords for analysing student satisfaction data. *29th Australasian Association for Engineering Education Conference 2018 (AAEE 2018)*, 132.

Cunningham-Nelson, S., Baktashmotlagh, M., & Boles, W. (2019). Visualizing Student Opinion Through Text Analysis. *IEEE Transactions on Education*, *62*(4), 305–311. https://doi.org/10.1109/TE.2019.2924385

De Witte, K., & Rogge, N. (2011). Accounting for exogenous influences in performance evaluations of teachers. *Economics of Education Review*, *30*(4), 641–653. https://doi.org/10.1016/j.econedurev.2011.02.002

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in *R*. *Journal of Statistical Software*, *25*(5). https://doi.org/10.18637/jss.v025.i05

Finch, W. H., Hernández Finch, M. E., McIntosh, C. E., & Braun, C. (2018). The use of topic modeling with latent Dirichlet analysis with open-ended survey items. *Translational Issues in Psychological Science*, *4*(4), 403–424. https://doi.org/10.1037/tps0000173

Gottipati, S., Shankararaman, V., & Gan, S. (2017). A conceptual framework for analyzing students' feedback. *2017 IEEE Frontiers in Education Conference (FIE)*, 1–8. https://doi.org/10.1109/FIE.2017.8190703

Gottipati, S., Shankararaman, V., & Lin, J. (2018a). Latent Dirichlet Allocation for textual student feedback analysis. *Proceedings of the 26th International Conference on Computers in Education ICCE 2018: Manila, Philippines, November 28-30*, 220–227.

Gottipati, S., Shankararaman, V., & Lin, J. R. (2018b). Text analytics approach to extract course improvement suggestions from students' feedback. *Research and Practice in Technology Enhanced Learning*, *13*(1), 6. https://doi.org/10.1186/s41039-018-0073-0

Grebennikov, L., & Shah, M. (2013). Student voice: using qualitative feedback from students to enhance their university experience. *Teaching in Higher Education*, *18*(6), 606–618. https://doi.org/10.1080/13562517.2013.774353

Grün, B., & Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(13), 1–30. https://doi.org/10.18637/jss.v040.i13

Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Information Processing & Management*, *54*(6), 1292–1307. https://doi.org/10.1016/j.ipm.2018.05.006

Hew, K. F., Hu, X., Qiao, C., & Tang, Y. (2020). What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers & Education*, *145*, 103724. https://doi.org/10.1016/j.compedu.2019.103724

Hoel, A., & Dahl, T. I. (2019). Why bother? Student motivation to participate in student evaluations of teaching. *Assessment & Evaluation in Higher Education*, *44*(3), 361–378. https://doi.org/10.1080/02602938.2018.1511969

Ibrahim, Z. M., Bader-El-Den, M., & Cocea, M. (2019). Mining Unit Feedback to Explore Students' Learning Experiences. In A. Lotfi, H. Bouchachia, A. Gegov, C. Langensiepen, & M. McGinnity (Eds.), *Advances in Computational Intelligence Systems* (pp. 339–350). Springer International Publishing. https://doi.org/10.1007/978-3-319-97982-3_28

Jacobi, C., Atteveldt, W. van, & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*(1), 89–106. https://doi.org/10.1080/21670811.2015.1093271

Jena, R. K. (2019). Sentiment mining in a collaborative learning environment: capitalising on big data. *Behaviour & Information Technology*, *38*(9), 986–1001. https://doi.org/10.1080/0144929X.2019.1625440

Joffe, H., & Yardley, L. (2004). Content and thematic analysis. *Research Methods for Clinical and Health Psychology*, *56*, 68.

Kember, D., Leung, D. Y. P., & Kwan, K. P. (2002). Does the Use of Student Feedback Questionnaires Improve the Overall Quality of Teaching? *Assessment & Evaluation in Higher Education*, *27*(5), 411–425. https://doi.org/10.1080/0260293022000009294

Klein, L. F., Eisenstein, J., & Sun, I. (2015). Exploratory Thematic Analysis for Digitized Archival Collections. *Digital Scholarship in the Humanities*, *30*(suppl 1), i130–i141. https://doi.org/10.1093/llc/fqv052

Langan, A. M., Scott, N., Partington, S., & Oczujda, A. (2017). Coherence between text comments and the quantitative ratings in the UK's National Student Survey. *Journal of Further and Higher Education*, *41*(1), 16–29. https://doi.org/10.1080/0309877X.2014.1000281

Lowenthal, P., Bauer, C., & Chen, K.-Z. (2015). Student Perceptions of Online Learning: An Analysis of Online Course Evaluations. *American Journal of Distance Education*, *29*(2), 85–97. https://doi.org/10.1080/08923647.2015.1023621

Marsh, H. W. (1987). Students' evaluations of University teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, *11*(3), 253–388. https://doi.org/10.1016/0883-0355(87)90001-2

Marsh, H. W. (2001). Distinguishing Between Good (Useful) and Bad Workloads on Students' Evaluations of Teaching. *American Educational Research Journal*, *38*(1), 183–212. https://doi.org/10.3102/00028312038001183

Marsh, H. W., & Roche, L. A. (2000). 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), 202–228.

Maxwell, J. (1992). Understanding and Validity in Qualitative Research. *Harvard Educational Review*, *62*(3), 279–301. https://doi.org/10.17763/haer.62.3.8323320856251826

McDonald, J., Moskal, A. C. M., Goodchild, A., Stein, S., & Terry, S. (2019). Advancing text-analysis to tap into the student voice: a proof-of-concept study. *Assessment & Evaluation in Higher Education*, 1–11. https://doi.org/10.1080/02602938.2019.1614524

McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–23. https://doi.org/10.1145/3359174

Meng Tan, T., & Chye Koh, H. (1997). Empirical investigation of the factors affecting SET results. *International Journal of Educational Management*, *11*(4), 170–178. https://doi.org/10.1108/09513549710186272

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–272.

Nimala, K., & Jebakumar, R. (2019). Sentiment topic emotion model on students feedback for educational benefits and practices. *Behaviour & Information Technology*, *0*(0), 1–9. https://doi.org/10.1080/0144929X.2019.1687756

Nitin, G. I., Swapna, G., & Shankararaman, V. (2015). Analyzing educational comments for topics and sentiments: A text analytics approach. *2015 IEEE Frontiers in Education Conference (FIE)*, 1–9. https://doi.org/10.1109/FIE.2015.7344296

Onan, A. (2019). Mining opinions from instructor evaluation reviews: A deep learning approach. *Computer Applications in Engineering Education*, *n/a*(n/a). https://doi.org/10.1002/cae.22179

Pan, D., Tan, G. S. H., Ragupathi, K., Booluck, K., Roop, R., & Ip, Y. K. (2009). Profiling Teacher/Teaching Using Descriptors Derived from Qualitative Feedback: Formative and Summative Applications. *Research in Higher Education*, *50*(1), 73–100. https://doi.org/10.1007/s11162-008-9109-4

Peng, X., & Xu, Q. (2020). Investigating learners' behaviors and discourse content in MOOC course reviews. *Computers & Education*, *143*, 103673. https://doi.org/10.1016/j.compedu.2019.103673

Poulos, A., & Mahony, M. J. (2008). Effectiveness of feedback: the students' perspective. *Assessment & Evaluation in Higher Education*, *33*(2), 143–154. https://doi.org/10.1080/02602930601127869

Pyasi, S., Gottipati, S., & Shankararaman, V. (2018). SUFAT - An Analytics Tool for Gaining Insights from Student Feedback Comments. *2018 IEEE Frontiers in Education Conference (FIE)*, 1–9. https://doi.org/10.1109/FIE.2018.8658457

Remedios, R., & Lieberman, D. A. (2008). I liked your course because you taught me well: the influence of grades, workload, expectations and goals on students' evaluations of teaching. *British Educational Research Journal*, *34*(1), 91–115. https://doi.org/10.1080/01411920701492043

Roberts, M. E., Stewart, B. M., & Tingley, D. (2018). *stm: R Package for Structural Topic Models*. http://www.structuraltopicmodel.com

Santhanam, E., Lynch, B., & Jones, J. (2018). Making sense of student feedback using text analysis – adapting and expanding a common lexicon. *Quality Assurance in Education*, *26*(1), 60–69. https://doi.org/10.1108/QAE-11-2016-0062

Sengkey, D. F., Jacobus, A., & Manoppo, F. J. (2019). Implementing Support Vector Machine Sentiment Analysis to Students' Opinion toward Lecturer in an Indonesian Public University. *Journal of Sustainable Engineering: Proceedings Series*, *1*(2), 194–198. https://doi.org/10.35793/joseps.v1i2.27

Shah, M., & Pabel, A. (2019). Making the student voice count: using qualitative student feedback to enhance the student experience. *Journal of Applied Research in Higher Education*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/JARHE-02-2019-0030

Sievert, C., & Shirley, K. (2015). *LDAvis: Interactive Visualization of Topic Models*. https://CRAN.R-project.org/package=LDAvis

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Sage Publications.

Spooren, P. (2010). On the credibility of the judge: A cross-classified multilevel analysis on students' evaluation of teaching. *Studies in Educational Evaluation*, *36*(4), 121–131. https://doi.org/10.1016/j.stueduc.2011.02.001

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the Validity of Student Evaluation of Teaching: The State of the Art. *Review of Educational Research*, *83*(4), 598–642. https://doi.org/10.3102/0034654313496870

Spooren, P., & Van Loon, F. (2012). Who Participates (not)? A Non-Response Analysis on Students' Evaluations of Teaching. *Procedia - Social and Behavioral Sciences*, *69*, 990–996. https://doi.org/10.1016/j.sbspro.2012.12.025

Srinivas, S., & Rajendran, S. (2019). Topic-based knowledge mining of online student reviews for strategic planning in universities. *Computers & Industrial Engineering*, *128*, 974–984. https://doi.org/10.1016/j.cie.2018.06.034

Stewart, M. (2015). The language of praise and criticism in a student evaluation survey. *Studies in Educational Evaluation*, *45*, 1–9. https://doi.org/10.1016/j.stueduc.2015.01.004

Stupans, I., McGuren, T., & Babey, A. M. (2016). Student Evaluation of Teaching: A Study Exploring Student Rating Instrument Free-form Text Comments. *Innovative Higher Education*, *41*(1), 33–42. https://doi.org/10.1007/s10755-015-9328-5

Ting, K. (2000). A Multilevel Perspective on Student Ratings of Instruction: Lessons from the Chinese Experience. *Research in Higher Education*, *41*(5), 637–661. https://doi.org/10.1023/A:1007075516271

Unankard, S., & Nadee, W. (2020). Topic Detection for Online Course Feedback Using LDA. In E. Popescu, T. Hao, T.-C. Hsu, H. Xie, M. Temperini, & W. Chen (Eds.), *Emerging Technologies for Education* (pp. 133–142). Springer International Publishing. https://doi.org/10.1007/978-3-030-38778-5_16

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*, 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

Wachtel, H. K. (1998). Student Evaluation of College Teaching Effectiveness: a brief review. *Assessment & Evaluation in Higher Education*, *23*(2), 191–212. https://doi.org/10.1080/0260293980230207

Wallace, S. L., Lewis, A. K., & Allen, M. D. (2019). The State of the Literature on Student Evaluations of Teaching and an Exploratory Analysis of Written Comments: Who Benefits Most? *College Teaching, 67*(1), 1–14. https://doi.org/10.1080/87567555.2018.1483317

Figure captions

Figure 1. The data-analysis process for responses to open-ended questions on student feedback surveys.

Figure 2. A simplified example of the data used in the estimation of multilevel models.

Figure 3. Illustration of the data structure. *Si* = student feedback *i* given to course *Cj* organised by study programme *Pk*.

Figure 4. The percentage shares of feedbacks assigned to topics

Appendices

A. Student feedback questionnaires used in data collection

[Place student feedback questionnaire 2016-2017 approximately here]

[Place student feedback questionnaire 2017-2018 approximately here]

B. A summary of codes assigned to the feedbacks during the thematic analysis process

Table B1. Codes assigned to the feedbacks during the thematic analysis for each topic.

Topic	Positive codes	Negative codes
Topic 1	*Total: 27* interesting (6) effective teaching (5) delivery of concepts (3) aids understanding (2) engaging (2) real-life applications (3) course design (1) course arrangements (1) lecture notes (1) clarity (1) willing to help (1) effective use of examples (1)	*Total: 26* course design (6) time management (4) pace of teaching (3) ineffective lecturing (2) ineffective slides (2) not interesting (1) unhelpful (1) delivery of concepts (1) problems in assessment (1) ineffective use of examples (1) ineffective use of concepts (1) unclear (1) uninteresting (1) problems with tutorials (1)
Topic 2	*Total: 6* interesting (4) engaging (1) effective teaching (1)	*Total: 72* unhelpful (7) problems with assessments (6) staff (6) pace of teaching (6) ineffective lecturing (5) disorganised (5) time management (5) course arrangements (5) course design (4) not interesting (4) unclear (4) problems with tutorials (4) ineffective slides (3) poor explanation (2) not enough real-life applications (2) problems with assessment (1) disorganised problems with assessments (1) problems with readings (1) not detailed enough (1)

Topic 3	*Total: 13* time management (2) interesting (2) assessment (2) ability to explain (1) approachable (1) knowledgeable (1) friendly (1) effective use of examples (1) good lecture notes (1) clarity (1)	*Total: 35* time management (6) unclear (4) not detailed enough (4) problems with assessments (3) pace of teaching (3) poor explanation (2) difficulty in understanding (2) problem with readings (2) disorganised (2) uninteresting (2) simplify explanations/concepts/terms (1) ineffective slides (1) ineffective use of examples (1) unhelpful (1) staff quality (1)
Topic 4	*Total: 5* aids understanding (1) willing to help (1) effective exercises (1) course structure (1) interesting (1)	*Total: 36* time management (8) unclear (5) problems with assessments (4) poor explanation (2) ineffective notes (2) unhelpful (2) ineffective lecturing (2) ineffective use of examples (2) not detailed enough (1) course structure (1) unprepared (1) ineffective exercises (1) practical theory links (1) disorganised (1) pace of teaching (1) difficulty in understanding (1) ineffective slides (1)
Topic 5	*Total: 20* interesting (6) effective use of examples (3) effective teaching (2) informative (1) ability to explain (1) real-life applications (2) approachable (1) willing to help (1) encouraging (1) delivery of concepts (1) engaging (1)	*Total: 47* difficulty in understanding (5) pace of teaching (4) ineffective lecturing (4) time management (4) ineffective use of concepts (4) unclear (3) ineffective use of examples (3) not detailed enough (2) simplify explanations/concepts/terms (2) not interesting (2) ineffective slides (2) relevance (2) poor questioning (2) ineffective notes (1) problem with readings (1) disorganised (1) not enough real-life applications (1) poor explanation (1) unhelpful (1) not detailed enough (1) problems with tutorials (1)

Topic 6	*Total: 37* effective teaching (8) interesting (8) teaching methods (5) knowledgeable (3) course structure (2) relevance (2) methods (1) ability to explain (1) clarity (1) aids understanding (1) delivery of concepts (1) real-life applications (1) effective questioning (1) effective use of examples (1) good lecture notes (1)	*Total: 6* time management (2) poor questioning (1) problems with assessments (1) not detailed enough (1) ineffective use of examples (1)

C. Descriptive statistics and estimation results of the multilevel models

Table C1. Descriptive statistics and pairwise Spearman correlation coefficients (2016-17 sample)

	M	SD	N	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1. Topic 1 %	16.40	3.42	2323	1									
2. Topic 2 %	17.79	4.44	2323	-.246*	1								
3. Topic 3 %	16.91	4.07	2323	-.295*	-.119*	1							
4. Topic 4 %	17.02	3.71	2323	-.127*	-.156*	-.212*	1						
5. Topic 5 %	16.69	3.44	2323	-.079*	-.253*	-.251*	-.180*	1					
6. Topic 6 %	15.19	3.22	2323	.007	-.218*	-.043	-.251*	-.061*	1				
7. Motivation	3.52	.91	2323	-.018	-.060*	.001	-.051*	-.004	.120*	1			
8. Perceptions of teaching	3.42	1.08	2316	.075*	-.153*	.003	-.057*	-.018	.309*	.481*	1		
9. Perceived workload	3.30	.85	1661	-.063*	.033	.092*	-.061*	.001	-.068*	.057*	-.045	1	
10. Perceived learning	3.42	1.13	2318	.041	-.142*	.018	-.050*	-.020	.253*	.614*	.702*	-.020	1

* p<0.05

Table C2. Descriptive statistics and pairwise Spearman correlation coefficients (2017-18 sample)

	M	SD	N	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. Topic 1 %	16.85	3.71	3496	1								
2. Topic 2 %	15.89	3.83	3496	-.220*	1							
3. Topic 3 %	16.28	3.71	3496	-.246*	-.053*	1						
4. Topic 4 %	16.49	3.73	3496	-.123*	-.072*	-.184*	1					
5. Topic 5 %	16.68	3.60	3496	-.078*	-.150*	-.179*	-.087*	1				
6. Topic 6 %	17.82	4.27	3496	-.111*	-.236*	-.103*	-.313*	-.236*	1			
7. Motivation	3.58	.88	3493	-.007	-.100*	.008	-.051*	-.100*	.227*	1		
8. Perception of implementation	3.48	1.01	3487	.071*	-.160*	.022	-.118*	-.094*	.326*	0.544*	1	
9. Perceived workload	3.29	.82	3478	-.091*	.007	.123*	.032	-.016	-.097*	0.113*	-0.041*	1

* p<0.05

Table C3. Estimated parameters of the three-level random intercept models predicting topic probabilities (2016-17 sample)

VARIABLES	Topic 1 (%)	Topic 2 (%)	Topic 3 (%)	Topic 4 (%)	Topic 5 (%)	Topic 6 (%)
Fixed effects						
Intercept	16.894**	17.361**	16.771**	16.853**	16.816**	15.253**
	(0.186)	(0.218)	(0.253)	(0.301)	(0.125)	(0.160)
Motivation	-0.169	0.401**	-0.0258	-0.0587	0.152	-0.316**
	(0.116)	(0.128)	(0.126)	(0.126)	(0.120)	(0.104)
Perceived workload	-0.0570	0.208	0.177	-0.248*	-0.0407	-0.128
	(0.100)	(0.110)	(0.109)	(0.107)	(0.103)	(0.089)
Perceptions of teaching	0.138	-0.498**	0.104	-0.225	-0.159	0.760**
	(0.114)	(0.125)	(0.125)	(0.121)	(0.117)	(0.101)
Perceived learning	0.0801	-0.534**	0.0770	0.0173	-0.045	0.354**
	(0.116)	(0.128)	(0.125)	(0.126)	(0.120)	(0.104)
Random effects						
var_v	0.107	0.212	0.219	0.582	0.000	0.106
	(0.139)	(0.160)	(0.230)	(0.344)	(0.000)	(0.089)
var_u	1.522	1.056	3.192	0.563	1.049	0.550
	(0.328)	(0.305)	(0.596)	(0.179)	(0.256)	(0.171)
var_e	10.036	12.372	11.470	12.101	10.883	8.229
	(0.366)	(0.452)	(0.422)	(0.434)	(0.395)	(0.299)
Fit statistics						
-2LL	8 639	8 945	8911	8 867	8 734	8 256
LR test chi-square (df = 4)	3 399**	4 020**	3 755**	3 456**	3 509**	3 572**
R^2	0.016	0.056	0.009	0.018	0.000	0.118
Observations	1 653	1 653	1 652	1 650	1 653	1 653
N. of courses	150	150	150	150	150	150
N. of study programmes	9	9	9	9	9	9

Note: The LR test compares the intercept-only model with the corresponding random intercept model; R^2 is Level 1 (student level) explained proportion of variance (Snijders and Bosker, 1999), *p<0.05, **p<0.01

Table C4. Estimated parameters of the three-level random intercept models predicting topic probabilities (2017-18 sample)

VARIABLES	Topic 1 (%)	Topic 2 (%)	Topic 3 (%)	Topic 4 (%)	Topic 5 (%)	Topic 6 (%)
Fixed effects						
Intercept	16.820**	15.733**	16.578**	16.506**	16.641**	17.712**
	(0.196)	(0.141)	(0.296)	(0.225)	(0.0898)	(0.228)
Motivation	-0.127	0.00442	-0.250**	0.267**	-0.209**	0.324**
	(0.0828)	(0.0817)	(0.0801)	(0.0819)	(0.0822)	(0.0919)
Perceived workload	-0.263**	-0.00836	0.368**	0.128	-0.00255	-0.324**
	(0.0762)	(0.0748)	(0.0741)	(0.0751)	(0.0752)	(0.0843)
Perceptions of	0.341**	-0.716**	0.127	-0.591**	-0.219**	1.165**
implementation	(0.0739)	(0.0727)	(0.0718)	(0.0729)	(0.0731)	(0.0817)
Random effects						
var_v	0.242	0.108	0.653	0.396	0.000	0.382
	(0.161)	(0.086)	(0.394)	(0.216)	(0.000)	(0.215)
var_u	1.522	0.759	2.670	0.892	0.830	1.140
	(0.240)	(0.135)	(0.369)	(0.160)	(0.156)	(0.219)
var_e	11.403	11.326	10.482	11.337	11.508	14.268
	(0.283)	(0.278)	(0.261)	(0.280)	(0.293)	(0.353)
Fit statistics						
-2LL	18 522	18 394	18 339	18 438	18 470	19 246

LR test chi-square (df = 3)	163**	261**	145**	187**	158**	484**
R^2	0.030	0.053	0.057	0.049	0.010	0.146
Observations	3 469	3 465	3 470	3 468	3 469	3 470
Number of courses	243	243	243	243	243	243
No. of study programmes	10	10	10	10	10	10

Note: LR test compares intercept-only model with the corresponding random intercept model; R^2 is Level 1 (student level) explained proportion of variance (Snijders and Bosker, 1999), *p<0.05, **p<0.01