Toward Automatic Interpretation of Narrative Feedback in Competency-Based Portfolios

This article has been accepted for publication in IEEE Transactions on Learning Technologies. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TLT.2022.3159334

Towards automatic interpretation of narrative feedback in competency-based portfolios

Joyce M.W. Moonen - van Loon, Marjan Govaerts, Jeroen Donkers, Peter van Rosmalen

***Abstract*— Self-directed learning is generally considered a key competence in higher education. To enable self-directed learning, assessment practices increasingly embrace assessment for learning rather than assessment of learning, shifting the focus from grades and scores to provision of rich, narrative and personalized feedback. Students are expected to collect, interpret and give meaning to this feedback, in order to self-assess their progress and to formulate new, appropriate learning goals and strategies. However, interpretation of aggregated, longitudinal narrative feedback has been proven to be very challenging, cognitively demanding and time consuming. In this study, we therefore explored the applicability of existing, proven text mining techniques to support feedback interpretation. More specifically, we investigated whether it is possible to automatically generate meaningful information about prevailing topics and the emotional load of feedback provided in medical students' competence-based portfolios (N = 1500), taking into account the competence framework and the students' various performance levels. Our findings indicate that the text mining techniques topic modeling and sentiment analysis make it feasible to automatically unveil the two principal aspects of narrative feedback, namely the most relevant topics in the feedback and their sentiment. This study therefore takes a valuable first step towards the automatic, online support of students, who are tasked with meaningful interpretation of complex narrative data in their portfolio as they develop into self-directed life-long learners.**

***Index Terms*—Assessment for learning, E-portfolio, Learning analytics, Narrative feedback, Text mining**

I. INTRODUCTION

RECENT developments in higher education, and the movement towards competency-based education in particular, resulted in a shift towards assessment for learning, with an intrinsic focus on formative assessments and assessment embedded in students' learning [1-4]. As a consequence, assessment practices witnessed a change from an almost exclusive focus on quantitative assessment data, like grades or scores, towards provision of rich, narrative and personalized feedback [5]. As modern curricula emphasize development of self-directed learning, students are expected to engage proactively in feedback processes and use of feedback for learning and performance improvement. Students are thus asked to gather and give meaning to narrative assessment data and to self-assess their progress in order to formulate new learning goals and strategies, based on their interpretation of received feedback. In many curricula, students' self-regulated learning and its development is supported by a mentor and an e-portfolio that contains narrative feedback data and reports on work done [6].

Research findings, however, indicate that interpretation of aggregated narrative feedback that is collected over longer periods of time and across various settings can be challenging for students and mentors, as feedback is likely to be composed of multiple comments of varying characteristics, e.g. positive and negative valence, referring to different competency domains and containing various suggestions for improvement. Complexity will increase in case of multiple feedback providers and assessors for which some degree of divergence is often present. Feedback can thus contain conflicting information and be described in various ways [7]. Furthermore, feedback providers tend to use specific linguistic strategies to present a message, to nuance earlier comments [8] or to carefully express criticism [9]. As a consequence of this so-called hedging, the feedback message can be unclear. Research on workplace-based assessments also shows that messages from narrative feedback and quantitative feedback data (performance scores) may diverge, further hindering clear interpretation of feedback in the portfolio [10-13]. Consequently, analysis of longitudinal feedback is not only complex but also time consuming. In times of high workload this may increase the chances of incomplete or inaccurate interpretation of assessment data even further.

Given the increasing importance and volume of narrative assessment data for students, it is important to consider how to support students to use these data for their learning in an efficient and accurate way.

One solution to facilitate the interpretation of complex narrative feedback can be found in learning analytics. Learning analytics is aimed at optimizing the process of data measurement, data collection, data analysis and reporting of data about learners and their learning contexts for the purpose of understanding and optimizing learning [14]. By processing large amounts of data automatically, many possibilities of interpretation and evaluation of assessment data in a learner's portfolio open up. Text mining techniques, aimed at supporting interpretation of narrative data, may not only save a lot of valuable time but may also enhance assessment quality by supporting interpretation of complex assessment data in the e-portfolios. They can be used to describe a student's strengths and weaknesses, provide a better understanding of a feedback provider's intentions and the role of context on student performance [15]. Text mining is widespread, and many different techniques and approaches are available.

Two aspects of narrative text that might help students to unravel their feedback is to help unveil the most important topics discussed in the feedback and the sentiment in which the feedback provider discussed them. There are so-called text mining techniques that support the search for underlying topics

in a large dataset (topic modeling) and that extract the underlying opinions, feelings or emotional load, i.e. level of positivity, of the writers of the texts (sentiment analysis). These two proven text mining techniques, that come in various forms, are commonly applied for the analysis of, for example, twitter-feeds on a certain topic, customer reviews on a product or service, public opinions on political questions or candidates [16-19]. However, use of text mining techniques for analysis of narrative data in assessment and e-portfolios in particular seems underexplored. Müller and Rebholz [20] describe an approach for automatic assessment of e-portfolios in Media Education and Management. It uses descriptive statistical tools (e.g. text lengths, number of links, word clouds or the appearance of typical concepts in the portfolio) and machine learning techniques to create topics that can be found in the narratives in the portfolio, using techniques similar to that for automatically grading essays. The underlying sentiment of feedback providers, however, is not addressed in this study. Ferreira-Mello et al. [21] state in their review study that specifically topic extracting and opinion determining tools and their applications on non-English datasets are important areas to investigate in text mining on educational data.

The goal of this study therefore is to explore the usability of learning analytics and take a first step towards the interpretation of a large and complex data set containing narrative feedback collected in a competency-based e-portfolio on the performance of students. More specifically, we investigate whether proven text mining techniques topic modeling and sentiment analysis are able to generate meaningful information about main feedback topics and the valence of feedback (i.e. emotional load) for each of these topics, taking the underlying competency framework and students' varying performance levels into account.

II. Setting

This study was conducted at Maastricht University, the Netherlands, master's programme in Medicine. The programme comprises three years of clinical clerkships in an academic hospital and affiliated teaching hospitals. The curriculum is designed according to the principles of competency-based education and assessment using the CanMEDS competencies as an overarching framework [6]. The assessment programme is supported by a web-based portfolio system in which students collect assessment data that are to be used for self-assessment and reflection on learning and development in each of the competency domains [22]. Assessment data consist of quantitative data but mainly also qualitative (narrative) feedback on performance in workplace settings. Assessment forms invite feedback providers to specify feedback for each of the competency domains and to clearly distinguish between strengths and weaknesses (i.e. suggestions for improvement). Based on this, at three different points in time (after two 9-week clerkships (T1), at the end of all five clerkships (T2) and at the end of the three year master's programme, including participations (T3) respectively), students receive a summative assessment on their progress, resulting in a qualification reflecting the student's performance level against performance standards (below, at or above expected level).

In the next paragraphs, we describe the various steps in our research method, summarized in the diagram presented in Fig. 1. All algorithms, techniques and statistical analyses in this study were performed using R [23].

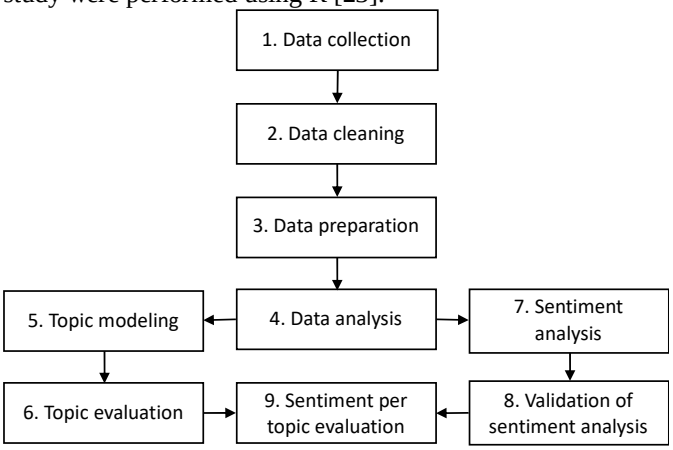

Fig. 1. Study workflow

III. Data

A. Data Collection

For this study, we used data collected in consenting students' e-portfolios between January 2013 and June 2019. Students explicitly gave their consent to use the data in their portfolio anonymously for the purpose of academic research. The narrative feedback was mainly written in Dutch. The complete database contained the portfolio data of 1,516 students and consists of 288,312 assessment forms filled in by 6,394 feedback providers and assessors. Given the homogeneity of assessment in the clerkships and deviations in the participations, we included all workplace-based assessments (WBA) collected during the clerkships (in periods T1 and T2).

The education and e-portfolio are competency-based. Therefore, the database was split into datasets, each dataset representing data related to a specific CanMEDS competency. For practical reasons, we only report results for the datasets on competencies Communicator and Professional, since experienced portfolio assessors stated that these contain the richest data.

B. Data Cleaning

The text in the two resulting datasets was cleaned by removing html tags, trimming white spaces, replacing special characters and removing numbers using R package textclean [24]. All words in the datasets were automatically compared to the Dutch dictionary in the hunspell package [25]. Of all deviating words, 3,646 words were clearly misspelled and subsequently corrected, e.g. "feedbakc", "iniaitief" instead of "initiatief" (initiative). A total of 1,191 words were added to the dictionary mainly involving words that are quite common in feedback in medical education, e.g. "aanpakker" (go-getter), "coachbaar" (coachable).

C. Data Preparation

For each dataset containing information of a specific competency, we were interested in the narrative feedback for every combination of performance level (i.e. below, at or above expected level) and type of feedback (i.e. strength or weakness). This led to six non-overlapping, narrative feedback documents, according to the usual terminology in automatic text analysis, for each competency. In our study, a *document* is defined as the collection of all cleaned narrative feedback of a certain type of feedback retrieved from the portfolios of all students that were assessed at a certain performance level. Feedback comments can consist of many sentences, each of which may address a different topic. Therefore, all feedback comments were split into sentences.

D. Data Analysis

For each competency and document, Table 1 presents the number of students included and the number of assessors that provided the feedback. It shows how many workplace-based assessments (WBA) were added in the first (before T1) or second (between T1 and T2) study period for that particular document, and the average score (scale 1-5) on those assessments for that competency. Note that assessments that contained narratives in the strength and in the weakness text area for a specific competency, were counted in both documents of the corresponding performance level and competency.

IV. Topic Modeling

Topic modeling is a method for classifying text documents that is able to discover natural groups of words, so-called topics, even when it is unclear what the underlying subjects of the texts are. The basic principles of topic modeling are that each document consists of a mixture of topics, and each topic consists of a mixture of words. We used Latent Dirichlet Allocation (LDA) [26, 27] to determine the words that belong to a specific topic and the mixture of topics that described each document simultaneously. In this study, we applied functions from the R package topicmodels [28] to the dataset to extract topics from the set of nouns and verbs in each document.

The workflow of the topic modeling method that was applied on the datasets, is presented in Fig. 2. The method required an optimal number of topics, *k*, to search for. The optimal number was initially unknown. Therefore, we estimated the range of topics (step 5.1) using R package ldatuning [29, 30]. For competency Professional, we estimated that the optimal number of topics was between 2 and 6, for Communicator between 2 and 5. Setting *k* equal to 5, we first ran an unsupervised LDA (step 5.2), which means that only the content in the dataset determined the outcome of the algorithm, without any intervention. LDA output a probability β per word per topic, which is the probability that the word belongs to that topic. The words with the highest β values in a topic gave in conjunction an indication of the subject of this topic.

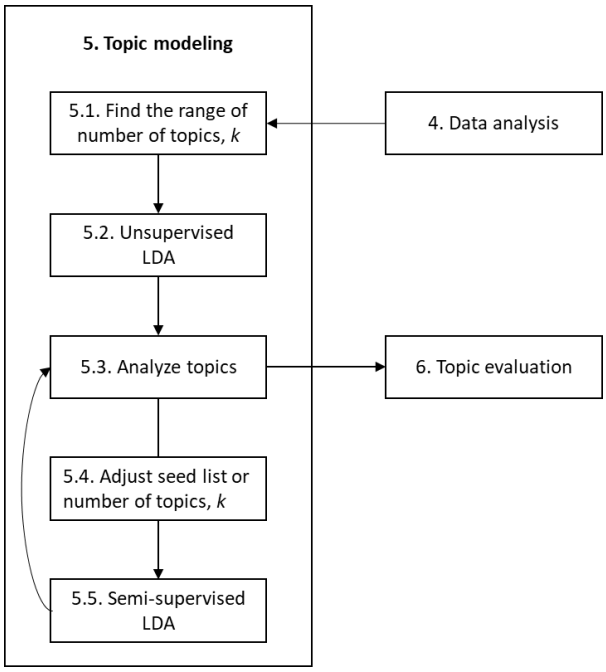

Fig. 2. Topic modeling workflow

Identified topics were analyzed in step 5.3, in which author 1 and author 2 investigated whether the ten words with the highest β value in a topic formed a cohesive, easily human-interpretable subject, and whether this subject differed from the

TABLE I
INFORMATION ON THE DOCUMENTS PER COMPETENCY

Competency	Document Level	Type	#students	#assessors	#WBA T1	#WBA T2	Average score
Professional	Below	Strength	11	36	53	17	3.91
Professional	Below	Weakness	11	28	54	7	3.52
Professional	At	Strength	776	2073	6282	8809	4.28
Professional	At	Weakness	767	1346	3554	3983	4.04
Professional	Above	Strength	602	1987	4310	9901	4.40
Professional	Above	Weakness	597	1147	1891	3654	4.22
Communicator	Below	Strength	12	94	181	68	3.60
Communicator	Below	Weakness	12	88	164	58	3.41
Communicator	At	Strength	752	2920	14567	14168	3.87
Communicator	At	Weakness	752	2637	11380	9242	3.66
Communicator	Above	Strength	657	2917	10436	19267	4.08
Communicator	Above	Weakness	657	2587	7176	11055	3.87

subjects in other topics. If there are not enough data in some of the documents or if many topics are used together in narrative feedback, the possible topics might get fused together in the unsupervised LDA. To overcome this, we can 'support' the algorithm by manually adding a few keywords to each found topic, so called seeds. By using seeding (step 5.4), the algorithm gave larger β values to relevant words occurring in the found topics, leading to more meaningful topics. These words were determined using exploratory analyses. For each individual document, we determined the word frequencies. For all documents together, we determined the co-occurrences of words, i.e. words that often occur together in sentences, using package udpipe [31]. The word frequencies and word co-occurrences (see Appendix A), i.e. co-occurrences within three words distance, per competency gave an indication to determine the seeds.

This process with seeding is called semi-supervised LDA [32, 33] (step 5.5). Again, the topics were analyzed leading to adjustments in *k* and/or the list of seed words based on earlier results, word frequencies and co-occurrences, to iteratively find the best fitting topic model for the dataset.

A. Topic Modeling Results

For competency Professional, the final set of four seeds for the semi-supervised topic modeling was (1) "initiatief", (2) "feedback", (3) "patiënt", (4) "vraag" (initiative, feedback, patient, ask), leading to five topics. For competency Communicator, the final set of three seeds was (1) "patiënt" & "communicatie", (2) "houding", (3) "informatie" (patient & communication, attitude, information), leading to four topics.

The LDA algorithm returned mixtures of words combined together in topics based on the sentences in the different documents. Fig. 3 presents the topics found and the ten words per topic with the highest β values within that topic. As heading, we present a manually assigned overarching subject per topic.

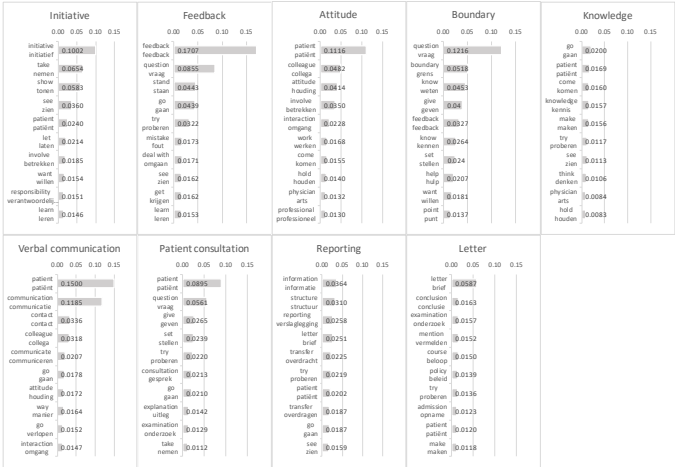

Fig. 3. Topics found in the documents for competencies Professional (above) and Communicator (below), with each ten words having the highest β values (x-axis), which is presented next to the word.

B. Topic Evaluation

During LDA, each sentence received a value γ per topic, which is the estimated proportion of words in a sentence that originate in a certain topic. In other words, γ is a value between 0 and 1 presenting the measure of fit for the sentence to a certain topic. The sum of all γ's per sentence over the topics is always equal to 1. In Appendix B and C, we present a few sentences per topic, retrieved from the different documents, with $\gamma > 0.97$ to show the typical sentences that are best fitting a certain topic.

Next, we analyzed how many sentences belonged to each topic to verify whether the majority of feedback was covered by the topics found by the LDA algorithm. We counted the sentences that belong to a topic with at least 50% probability, thus $\gamma > 0.5$, to assure that every sentence clearly belonged to at most one of the topics. Table 2 shows that 89% of the sentences in the dataset was related to one of the found topics for competency Professional and 93% for Communicator, presented in the second row by their overarching subjects. In every document all topics were covered, which is well explainable for the type of data used in this study. Whether a student performs poorly or very well, the topics of assessment were comparable. However, there were some subtle differences between the documents. We saw that for the competency Professional, strengths in performances were mainly described in relation to "Feedback" for those students who perform below expected level (28%), and mostly as related to "Attitude" for other students (24%). For weaknesses, feedback providers seemed to focus on topic "Knowledge" (26-34%). For competency Communicator, strengths in performances were mainly described as related to "Verbal communication" (32-37%) in all categories. For weaknesses, feedback providers seemed to focus on topic "Letter" (28-33%). The differences between the last three topics in students performing below expected level was rather small. For students that performed at or above expected level the main focus in constructive feedback

TABLE 2
NUMBER OF SENTENCES PER DOCUMENT, COMPETENCY AND TOPIC WITH $\gamma > 0.5$

Document		Professional						Communicator				
Level	**Type**	Initiative	Feedback	Attitude	Boundary	Knowledge	unassigned	Verbal communication	Patient consultation	Reporting	Letter	unassigned
Below	**Strength**	12	26	22	13	15	6	99	62	68	56	29
Below	**Weakness**	22	19	8	13	36	9	47	96	98	108	36
At	**Strength**	2966	4005	5024	3567	3117	1951	12989	8456	7888	6530	2428
At	**Weakness**	2260	1602	1093	1601	2752	1127	2770	8551	8141	10710	2224
Above	**Strength**	3153	3893	4989	3281	3211	1948	15420	9608	8212	6271	2630
Above	**Weakness**	1471	1158	796	1224	1852	760	2362	7744	6785	8530	1825

was on "Letter" (31-33%), followed by "Patient consultation" (26-28%) and "Reporting" (25%).

V. SENTIMENT ANALYSIS

Sentiment analysis is the computational study of opinions, sentiments and emotions expressed in text [34]. In this study, we applied the sentiment analysis techniques introduced by De Smedt and Daelemans [35], which are suited for the analysis of Dutch texts. As result, each sentence within a feedback comment received a polarity, which indicates the extent to which the sentence expresses a positive (maximum +1) or negative (minimum -1) sentiment, based on the words in the sentence and their co-occurrences.

After an initial analysis of the polarity of sentences from the narrative feedback in the students' portfolios, several context-specific deficiencies in the lexicon became apparent. Consider the following example: The sentence "Hij is erg behulpzaam" (He is very helpful) received a polarity of -0.35, implying that the sentence expresses a negative sentiment although it clearly is a positive statement on the performance of a medical student. However, the word "behulpzaam" (helpful) was not part of the lexicon. Therefore, in Dutch, the sentiment was determined only on the sentence "Hij is erg" (translated to "He is bad"), leading to a polarity of -0.35. In total, 148 sentiment-expressing words (mainly adjectives) in the dataset that were not included in the available lexicon of package pattern.nlp [36], were added to the lexicon with a polarity and subjectivity score. These scores were based either on the word's synonyms or on the translated word in English, as described in the article [35] and the already available lexicon extension [37].

We found that sentences written in imperative sense were not always determined as such by the algorithm, leading to polarities that were too positive. For example, whereas the statement "mag meer initiatief tonen" (may show more initiative) indicates that the student doesn't show enough initiative, the automatic polarity is 0.2, equal to sentence "you show more initiative". To overcome this, we adapted the algorithm for weaknesses to add the specific knowledge that the written narrative is an expression of behavior that is not perfect yet when the feedback started with words like "try", "may", "more" or "pay attention to". Hereby, we added some degree of context and the intended purpose of the feedback to the algorithm.

A. Validation of sentiment analysis

Because of the addition of context-specific adjectives in the lexicon, we needed to verify whether the accuracy of the sentiment analysis was still acceptable for the dataset at hand. Therefore, a random sample (N=593) of the available sentences was selected from the dataset, equally divided among the twelve documents. For each sentence, author 1 decided manually whether the statement was positive (polarity >= 0) or negative. These values were compared to the automatically determined polarity of each sentence.

The results of the polarity comparison of the sample of 593 sentences are presented in Fig. 4, in which the number of sentences is shown for each combination of manually and automatically estimated positive (≥ 0) and negative (< 0) polarity and various statistical values are determined.

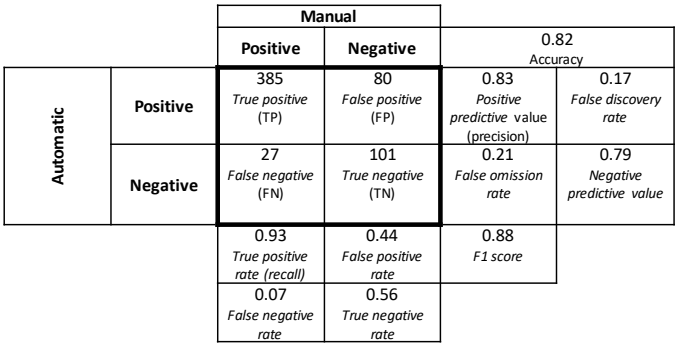

Fig. 4. Results of polarity comparison on a sample of 593 sentences

The accuracy of the sentiment analysis on the selected sample was equal to (385+101)/593 = 0.82, or 82%, which is the percentage of sentences of which the polarity is correctly estimated. The precision (or Positive Predictive Value) shows the percentage of sentences that were actually positive out of the set of sentences that were automatically estimated as positive. Precision, defined as TP/(TP+FP), was equal to 0.83.

We saw only a small difference between the positive and negative predictive value (precision). Recall is the percentage of sentences that are estimated as positive out of the set of all actual positive sentences, defined as TP/(TP+FN), was 0.93. Specificity (or True Negative Rate) is the percentage of sentences that were estimated as negative out of the set of actual negative sentences, defined as TN/(TN+FP), was 0.56. The F1-score, used to gauge the similarity between the manually and automatically estimated polarities, defined as 2 * (precision * recall) / (precision + recall), was 0.88. These results are comparable to earlier published results [35], where the accuracy is 82% (precision 0.80, recall 0.86, F1 0.83) for Dutch book reviews.

B. Results

The sentiment analysis was applied to each sentence to determine the polarity between -1 and +1. Table 3 presents the average and standard deviation of the polarity for all sentences within each document as well as the number of sentences (N). In general, we observed a larger average polarity for better performing students. Furthermore, the average polarity for statements added as strengths was higher than the statements added as weaknesses.

We determined the influence of the type of feedback (strength or weakness) and the level of performance on the emotional load (polarity) of the feedback, for both competencies. We concluded from the results of the two-way ANOVA that, with a p-value of 0.05, only the type of feedback (strength / weakness) had a significant effect on the polarity for competency Professional ($p < 2e\text{-}16$), and both the feedback type ($p < 2e\text{-}16$) and the performance level ($p = 7.26e\text{-}6$) had a significant effect on the polarity for competency Communicator. There was no interaction.

C. Sentiment per Topic Evaluation

In this study, we are interested in the emotional load of the feedback for each topic, given the competency-based e-portfolio and differences in types of feedback and performances of the students. The sentiment analysis assigned a polarity to each sentence. Table 4 presents the average polarity of all sentences per document that belonged to a topic with $\gamma > 0.5$, for both competencies. The topics are presented with their overarching subjects. We found that for students performing below the expected level, the most positive comments on average for competency Professional were written about topic "Feedback" as strengths, and for other students, the strengths were most positive in topic "Attitude". In all documents, the most negative comments expressed as weakness were described in topic "Feedback". For competency Communicator, the most positive strength comments for all students were found in topic "Verbal communication". The most negative comments were given in topic "Patient consultation" for all documents except for the strengths of students performing below expected level. For that group the most negative comments as strength were in topic "Letter".

TABLE 3
AVERAGE (STANDARD DEVIATION) OF THE POLARITY FOR ALL (N) SENTENCES PER COMBINATION OF PERFORMANCE LEVEL AND TYPE OF FEEDBACK, PER COMPETENCY.

	Professional			**Communicator**		
	Strength	**Weakness**	**All**	**Strength**	**Weakness**	**All**
Below	0.25 (0.28) N: 94	-0.05 (0.14) N: 107	0.09 (0.27) N: 201	0.32 (0.24) N: 314	-0.06 (0.15) N: 385	0.11 (0.27) N: 699
At	0.27 (0.25) N: 20630	-0.04 (0.16) N: 10435	0.16 (0.27) N: 31065	0.35 (0.24) N: 38291	-0.05 (0.15) N: 32396	0.17 (0.28) N: 70687
Above	0.26 (0.25) N: 20475	-0.04 (0.16) N: 7261	0.18 (0.27) N: 27736	0.36 (0.24) N: 42141	-0.04 (0.15) N: 27246	0.20 (0.29) N: 69387
All	0.26 (0.25) N: 41199	-0.04 (0.16) N: 17803	0.17 (0.27) N: 59002	0.35 (0.24) N: 80746	-0.04 (0.15) N: 60027	0.18 (0.28) N: 140773

TABLE 4
AVERAGE POLARITY PER DOCUMENT PER TOPIC

Document		**Professional**					**Communicator**			
Level	**Type**	Initiative	Feedback	Attitude	Boundary	Knowledge	Verbal communication	Patient consultation	Reporting	Letter
Below	**Strength**	0.18	0.35	0.28	0.18	0.16	0.38	0.29	0.29	0.24
Below	**Weakness**	-0.05	-0.08	-0.04	-0.07	-0.04	-0.06	-0.08	-0.07	-0.03
At	**Strength**	0.26	0.25	0.35	0.21	0.24	0.42	0.30	0.33	0.32
At	**Weakness**	-0.04	-0.07	-0.04	-0.05	-0.03	-0.04	-0.05	-0.05	-0.04
Above	**Strength**	0.26	0.23	0.34	0.20	0.25	0.42	0.30	0.33	0.32
Above	**Weakness**	-0.04	-0.05	-0.02	-0.04	-0.04	-0.03	-0.05	-0.04	-0.04

VI. Discussion

In this study, we explored the application and combination of topic modeling and sentiment analysis with extended lexicon on two datasets with narrative feedback in Dutch on the performance of medical students, collected from their e-portfolios within the competency-based master's programme in Medicine at Maastricht University, the Netherlands. Overall, the results indicate that the selected text mining techniques, topic modeling and sentiment analysis, lead to insights on the feedback, similar to the results in other research areas [18, 19]. However, it is important to view these results as a support for interpretation of feedback rather than a tool for automatic assessment, as also concluded by Müller and Rebholz [20]. In this study we took an important first step towards the delivery of useful and meaningful support and insight in the underlying opinions in narrative feedback to help individual students interpret the longitudinal feedback collected in their e-portfolio, which is a rather unexplored area in education research [21].

We found that we were able to automatically retrieve topics from the dataset that are recognizable and meaningful to humans, representing the underlying data and leading to more insight on the main subjects of the provided feedback. The sentences that are automatically assigned to the topics correspond to the overarching subject of the topic as interpreted by human experts, providing support for the usefulness of our topic modeling approach in interpretation of narrative, competency-based performance feedback.

Findings show that there are some differences in the number of sentences per topic for the different performance levels and types of feedback. Providing students with an overview of feedback quantity, gives them the valuable opportunity to request *additional* feedback on specific topics if necessary. Feedback providers tend to focus on different topics when commenting on strengths (attitude, verbal communication and contact with patients and colleagues, for example) compared to comments on weaknesses (knowledge and writing of reports and letters). Moreover, we found that the feedback is more positive for better performing students, which is significant for one of the competencies, and comments are significantly more positive for strengths than weaknesses. When observing the average polarities over the topics we noticed that for the weaknesses, the worse a student performs the more negative is the feedback on all topics. However, for the strengths, there is a clear difference between students performing below expected level and the others. Especially on the subject on dealing with feedback, these underperforming students receive more positive comments than the other students, whereas on knowledge, taking initiative and writing letters the comments are less positive. These findings are in line with previous research on supervisors' approach to describing various levels of trainee performance [9]. The study by Ginsburg and colleagues showed that supervisors focused on different aspects when describing problematic versus outstanding learners, and that aspects of performance may take on varying degrees of importance depending on the learner. We therefore feel that our findings from automated analysis of narrative feedback reflect authentic supervisor behaviors, further supporting our approach.

We were able to automatically determine the emotional load (polarity) for each Dutch feedback sentence with an accuracy of 82% using a context specific extended lexicon. Although the accuracy is comparable to that reported in earlier studies, we noticed a rather large number of false positives, meaning that the algorithm relatively often assigns a positive polarity to sentences that are manually labeled as negative. Further analysis of the statistics showed that, in this sample, all false positives were weaknesses. Possible reasons for this might be that the type of sentences (e.g. imperative tense) are not well enough discovered by the algorithm, or too many words in sentences are evaluated without polarity, due to still missing specific jargon in the lexicon. We therefore expect the differences in polarity between strengths and weaknesses to be even larger than presented in this study, since most comments with a manually assigned negative emotional load are provided as weakness that were not picked up by the algorithm.

There are some limitations to consider. First of all, the raw datasets contained a lot of typos. Therefore, an extensive pre-processing step was needed. For future studies on a similar dataset, this can be done automatically, but it should be (partly) reconsidered when applying the same technique on different competencies or research areas. Secondly, the available lexicon lacked content-specific wording for medical education. Many words were added and can be reused, but this step needs to be repeated for each specific context. Thirdly, in a lot of comments, feedback providers did not write full sentences and some feedback takes on a different meaning when added as a strength or as a weakness (e.g. "the way in which you show empathy"). Therefore, design of the portfolio should guide feedback providers in where and how to document narrative comments, but even more important, feedback providers should be trained in how to write clear and unambiguous, yet concise and meaningful feedback comments. Finally, it appeared to be a challenge to find a suitable sentiment analysis method for Dutch texts, more specifically texts in medical education. The used package seemed to be the best available at the moment and had a clear advantage of adding content-specific words to the lexicon, but also presented some limitations, especially in the parts-of-speech (POS) tagging for sentences written in imperative sense. Packages with more extensive POS taggers, however, lacked functionalities on polarity determination. Also, we noticed that the number of false positives is rather large, especially for weaknesses, leading to a low 'True negative rate'. Possible solutions to overcome this and further improve the accuracy, would be to improve the tagger of the used package, more extensive preprocessing to overcome the difficulties in the tagging, further extend the lexicon, or to translate the sentences automatically to English as there are many more available packages for sentiment analysis in English. The problem with this latter approach in this particular context might be that automatic translation asks for uploading portfolio data, leading to data security and privacy issues.

Even though the results in this study on applying proved text mining techniques on narrative portfolio data seem very

encouraging for the future, this is only the first step towards fully automated support on feedback interpretations for students. Some validations on the output of the applied techniques have been done in this study, but more research is needed to validate the results on these particular datasets by experts and to evaluate its added value to students and other portfolio users. The next step would be to support individual students by presenting clear (just-in-time) overviews of their specific strengths and weaknesses. The basics of the required algorithms will be similar, whether applied to a dataset on portfolios of many students or just one. However, the most benefit for individual students will be achieved when a student's specific strengths, weaknesses and gaps are presented by comparing their personal feedback to the intended learning outcomes as defined in the overarching competency framework. This will require a seeding in the topic modeling to capture the described competencies of the framework at a more granular level, and by further fine-tuning the lexicon of the sentiment analysis. Furthermore, an online integration with the e-portfolio leads to real time presentation of the personal narrative feedback, providing the opportunity for students to direct their learning, e.g. by specifically *asking* for tasks and feedback on their personal topics of improvement or the ones with a limited amount of feedback. In a current 18-month project, educationalists, IT-specialists, students and mentors work together to investigate and develop via a design based research approach ways to apply these techniques to support individual students optimally in their learning process and integrate the approach in the portfolio system. These results will not only be of added value for students, but also for their mentors when guiding the student's learning and for examiners that assess the portfolios of the students. Finally, future studies that compare the results of the sentiment analysis and thereby the underlying intention of the feedback giver, and the quantitative assessment data (numeric scores) collected on the same assessment tasks, might show differences and similarities in feedback and scoring, helping to identify gaps in the global analysis of performance [15] and aiding the discussion on the value of scores in performance and competence assessments.

In conclusion, we believe that this study is a valuable first step towards the automatic, online support of students, who are tasked with meaningful interpretation of complex narrative data in their portfolio as they develop into self-directed life-long learners.

Ethical approval: The conduct of the study is approved by the NVMO Ethical Review Board (NERB dossier number: 2019.5.1).

APPENDIX

A. *Word cooccurrence*

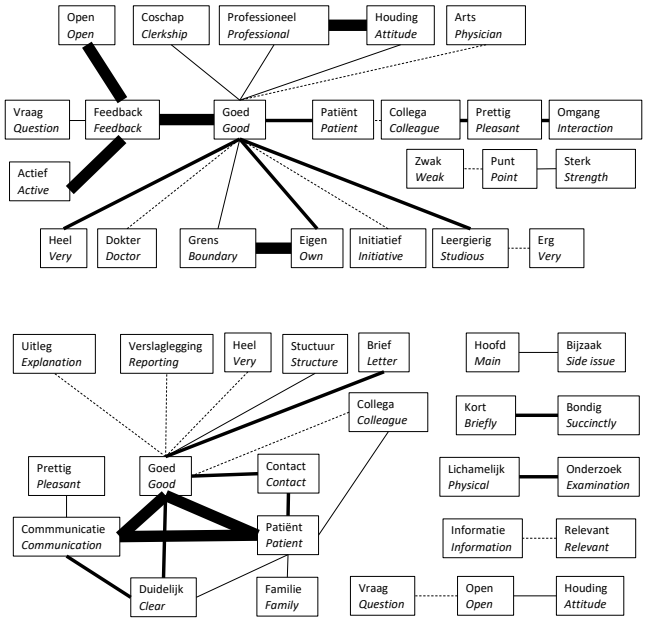

Fig. 5. Graphical presentation of the word co-occurrences within 3 words distance in the datasets Professional (upper images) and Communicator (lower images). Stronger lines indicate more co-occurrences.

B. *Selection of sentences for competency Professional*

Small selection of feedback sentences that belong to a particular topic, retrieved from various portfolios, for competency Professional, translated to English.

1: ***Initiative***

"I got to know the student as an enthusiastic and committed participant with a clear input and a pleasant, constructive attitude, for example during the simulated neighborhood team. I wish her the best of luck in completing her medical study"

"shows a lot of commitment and enthusiasm, after having previously been to Brunssum where the clinic was canceled due to circumstances, still came to Heerlen on his own initiative with public transport to follow the clinic"

"Try not to avoid too much at first. Direct involvement and offering yourself for work results more quickly in enthusiasm from both sides and thus more learning moments than waiting"

2: ***Feedback***

"Whenever possible, try to ask for feedback in advance, then I can also focus on certain things. Try to meet the assessor when the feedback form is filled in, so that you have it sooner and you can, for example, ask for an explanation if it's unclear"

"when make mistakes and get feedback on this, I take this well and try to learn from this to apply it next time and show that I have learned from what I was told"

3: ***Attitude***,

"As said, this student is worthy of a doctor; could function as a department doctor or in general practitioner training; has an eye for patient, context and team; is reliable and consistent in

his work and flexible in his efforts"

"involved with patients, wants to be in family conversations and keeps in touch; keeping an appropriate distance from the patient in a pleasant way and yet being compassionate, comes across as professional"

4: ***Boundary***

"Your open attitude; you actively ask for feedback and also give feedback in a pleasant way; you know your limits well and ask for help on time, or ask for someone to watch you and learn from that"

5: ***Knowledge***

"rising curve in recent weeks with regard to making differential diagnosis for the most common clinical pictures at the outpatient clinic; he also shows that he has gained the knowledge after further questioning"

"in the context of this case-based discussion (end-of-life decisions): well-developed insight into ethical aspects, good insight into ethical dilemmas, e.g. meeting euthanasia due care criteria, euthanasia in dementia"

"good medical knowledge, already at the start performed at expected level and with moderately complex problems, little guidance is required; solid internship, showing good learning curve, performed at a good level throughout the internship; combination of knowledge and communication skills are a good starting point for functioning as a doctor"

"not so much an improvement point, but something that could take you to the next level: keep looking critically at your own knowledge and skills and care, be aware of your strengths and weaknesses, and try to take yourself out of your comfort zone to become proficient in these weaker traits"

C. Selection of sentences for competency Communicator

Small selection of feedback sentences that belong to a particular topic, retrieved from various portfolios, for competency Communicator, translated to English.

1: ***Verbal communication***

"Communication with patient (regardless of age), parents, caregivers of patient, colleagues (nurse, desk clerk, interns, physician assistants and supervisors) is very good"

"I feel that communication with patients and family is natural, and I really enjoy listening to and helping people with their problem; communication with the GPs, other interns and other colleagues went very well, I felt at ease in the team"

2: ***Patient consultation***

"Gets to the same level as the patient; word usage adapted to patient; let the patient tell the story himself who at first does not dare or does not want to tell himself (by means of further questions, emotional reflection and encouragement); adequate brief anamnesis of a child with pain without omitting relevant things"

"Recognizes feelings in patient; gives the patient space to speak; uses earlier statements of the patient as a starting point to switch to another part of MSO, ensures a smooth transition"

"no explanation given prior to the physical examination, even though you say you are going to examine the patient from head to toe, she still doesn't know exactly what is going to happen and she lets you know non-verbally (by keeping her arms on her stomach) that she doesn't feel comfortable"

3: ***Reporting***

"Be careful not to stay in the background too much during meetings, visits, transfer moments; do realize that this is your first internship and that you have to grow in this; make a learning goal for the next internship; learn to present patient problems and to do more visits independently"

"neat extensive anamnesis, make this structured for yourself, try to interrupt the patient at a certain point to find out what you want to know; neat status management and therefore a neat letter"

"Please make sure that when sending or submitting files via email that the correct version of documents is used, otherwise it is a waste of your time and a pity about the missed feedback on the correct version of the end product"

4: ***Letter***

"All parts were neatly arranged: reason for admission, history, medication, anamnesis, clinical examination, lab, additional examination, discussion, conclusion; in your discussion you already show very accurate Dutch language in your problem description"

"try not to use abbreviations and make an ongoing story of the development; letter structure: reason for admission, history, additional examination (if relevant for GP), procedure, development, comments, medication, check-up"

REFERENCES

[1] D. Boud and R. Soler, "Sustainable assessment revisited," *Assessment & Evaluation in Higher Education,* vol. 41, no. 3, pp. 400-413, 2016.

[2] J. L. Hanson, A. A. Rosenberg, and J. L. Lane, "Narrative descriptions should replace grades and numerical ratings for clinical performance in medical education in the United States," *Front. Psychol.,* vol. 4, no. 668, 2013.

[3] L. W. Schuwirth and C. P. M. van der Vleuten, "Programmatic assessment: from assessment of learning to assessment for learning," *Med. Teach.,* vol. 33, no. 6, pp. 478-485, 2011.

[4] M. Taras, "Using Assessment for Learning and Learning from Assessment," *Assessment & Evaluation in Higher Education,* vol. 27, no. 6, pp. 501-510, 2002.

[5] D. J. Nicol and D. Macfarlane-Dick, "Formative assessment and self-regulated learning: A model and seven principles of good feedback practice," *Studies in higher education,* vol. 31, no. 2, pp. 199-218, 2006.

[6] E. W. Driessen, J. van Tartwijk, P. Teunissen, M. Govaerts, and C. P. M. van der Vleuten, "The use of programmatic assessment in the clinical workplace: A Maastricht case report," *Med. Teach.,* vol. 34, no. 3, pp. 226-231, 2012.

[7] P. Yeates, P. O'Neill, K. Mann, and K. Eva, "Seeing the Same Thing Differently: Mechanisms That Contribute to Assessor Differences in Directly-Observed Performance Assessments," *Advances in health science Education* vol. 18, no. 3, pp. 325-341, 2013, doi: 10.1007/s10459-012-9372-1.

[8] M. Bogo, C. Regehr, M. Woodford, J. Hughes, R. Power, and G. Regehr, "Beyond Competencies: Field Instructors' Descriptions of Student Performance," *Journal of Social Work Education*, vol. 42, no. 3, pp. 579-593, 2006.

[9] S. R. Ginsburg, J. McIlroy, O. Oulanova, K. Eva, and G. Regehr, "Toward Authentic Clinical Evaluation: Pitfalls in the Pursuit of Competency," *Acad. Med.*, vol. 85, no. 5, pp. 780-786, 2010.

[10] G. S. Cohen, P. Blumberg, N. C. Ryan, and P. L. Sullivan, "Do final grades reflect written qualitative evaluations of student performance?," *Teach. Learn. Med.*, vol. 5, no. 1, pp. 10-15, 1993.

[11] N. L. Dudek, M. B. Marks, and G. Regehr, "Failure to fail: the perspectives of clinical supervisors," *Acad. Med.*, vol. 80, no. 10, pp. S84-S87, 2005.

[12] M. Yepes-Rios, N. L. Dudek, R. Duboyce, J. Curtis, R. J. Allard, and L. Varpio, "The failure to fail underperforming trainees in health professions education: A BEME systematic review: BEME Guide No. 42," *Med. Teach.*, vol. 38, no. 11, pp. 1092-1099, 2016.

[13] S. R. Ginsburg, "Hidden in plain sight: the untapped potential of written assessment comments," Maastricht University, Maastricht, 2016.

[14] M. van der Schaaf *et al.*, "Improving workplace-based assessment and feedback by an E-portfolio enhanced with learning analytics," *Educational Technology Research and Development*, vol. 65, no. 2, pp. 359-380, 2017.

[15] T. M.-Y. Chan, S. Sebok-Syer, B. Thoma, A. Wise, J. Sherbino, and M. Pusic, "Learning Analytics in Medical Education Assessment: The Past, The Present, and The Future," *AEM Education and Training*, vol. 2, 2018, doi: 10.1002/aet2.10087.

[16] C. Gibbons, S. Richards, J. M. Valderas, and J. Campbell, "Supervised Machine Learning Algorithms Can Classify Open-Text Feedback of Doctor Performance With Human-Level Accuracy," *J. Med. Internet Res.*, vol. 19, no. 3, p. e65, 2017, doi: 10.2196/jmir.6533.

[17] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," *Proceedings of the Association for Computational Linguistics*, vol. 42, pp. 271-278, 2004.

[18] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," *Proceedings of the Association for Computational Linguistics*, vol. 43, pp. 115-124, 2005.

[19] B. Dahal, S. A. P. Kumar, and Z. Li, "Topic modeling and sentiment analysis of global climate change tweets," *Social Network Analysis and Mining*, vol. 9, no. 24, 2019, doi: https://doi.org/10.1007/s13278-019-0568-8.

[20] W. Müller, S. Rebholz, and P. Libbrecht, "Automatic Inspection of E-Portfolios for Improving Formative and Summative Assessment," in *Emerging Technologies for Education. SETE 2016. Lecture Notes in Computer Science*, vol. 10108, G. R. Wu TT., Huang YM., Xie H., Cao Y. Ed.: Springer, Cham, 2017.

[21] R. Ferreira-Mello, M. André, A. Pinheiro, E. Costa, and C. Romero, "Text mining in education," *Data Mining and Knowledge Discovery*, vol. 9, p. e1332, 2019.

[22] A. Oudkerk Pool, M. J. B. Govaerts, D. A. D. C. Jaarsma, and E. W. Driessen, "From Aggregation to Interpretation: How Assessors Judge Complex Data in a Competency-Based Portfolio," *Adv Health Sci Educ Theory Pract*, vol. 23, no. 2, pp. 275-287, 2018.

[23] *R: A language and environment for statistical computing*. (2019). R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: https://www.R-project.org/

[24] *textclean: Text Cleaning Tools*. (2018). Buffalo, New York. [Online]. Available: https://github.com/trinker/textclean

[25] *hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker*. (2018). [Online]. Available: https://CRAN.R-project.org/package=hunspell

[26] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993-1022, 2003.

[27] J. Silge and D. Robinson, "tidytext: Text Mining and Analysis Using Tidy Data Principles in R," *Journal of Open Source Software*, vol. 1, no. 3, 2016, doi: 10.21105/joss.00037.

[28] B. Grün and K. Hornik, "topicmodels: An R Package for Fitting Topic Models," *Journal of Statistical Software*, vol. 40, no. 13, pp. 1-30, 2011, doi: 10.18637/jss.v040.i13.

[29] *ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters*. (2020). [Online]. Available: https://CRAN.R-project.org/package=ldatuning

[30] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang, "A density-based method for adaptive lda model selection," *Neurocomputing — 16th European Symposium on Artificial Neural Networks 2008*, vol. 72, no. 7-9, pp. 1775-1781, 2009, doi: http://doi.org/10.1016/j.neucom.2008.06.011.

[31] *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*. (2019). [Online]. Available: https://CRAN.R-project.org/package=udpipe

[32] J. Jagarlamudi, H. Daumé III, and R. Udupa, "Incorporating Lexical Priors into Topic Models," *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pp. 204–213, 2012.

[33] V. Singh. "How we Changed Unsupervised LDA to Semi-Supervised GuidedLDA." https://www.freecodecamp.org/news/how-we-changed-unsupervised-lda-to-semi-supervised-guidedlda-e36a95f3a164/ (accessed.

[34] B. Liu, "Sentiment Analysis and Subjectivity," in *Handbook of Natural Language Processing*, N. Indurkhya and F. J. Damerau Eds., Second edition ed., 2010.

[35] T. De Smedt and W. Daelemans, ""Vreselijk mooi!" (terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives," in *LREC*, 2012.

[36] *pattern.nlp: R package to perform sentiment analysis for Dutch/French/English and Parts of Speech tagging for Dutch/French/English/German/Spanish/Italian*. (2016). [Online]. Available: https://github.com/bnosac/pattern.nlp

[37] [*Subjectivity lexicon for Dutch adjectives*. (2014). [Online]. Available: https://github.com/clips/pattern/blob/master/pattern/text/nl/nl-sentiment.xml

Joyce M.W. Moonen – van Loon was born in Heerlen, the Netherlands, in 1982. She received her M.Sc. degree in econometrics & operations research from Maastricht University, the Netherlands, in 2004 and the Ph.D. degree in operations research from Maastricht University, the Netherlands in 2009.

She worked as Project Leader and Senior ICT Developer to develop and implement e-portfolios in (medical) education programs using a programmatic assessment approach for the past 10 years. Currently, she works as an Assistant Professor in the Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences at Maastricht University. She is a member of the taskforce Instructional Design and E-learning. Her research focuses on how to use technology and mathematics to support the learning process of students and the supervising and assessing processes of teachers.

Marjan (M.J.B.) Govaerts was born in Beek (Limburg), the Netherlands in 1959. She received her MD degree from Maastricht University (the Netherlands) in 1982 and the Ph.D. degree in medical education (with distinction) from Maastricht University in 2011.

She is an associate professor in the Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences (FHML) at Maastricht University. She is the current chair of the taskforce on student assessment, which is the group responsible for various aspects of quality control and improvement of assessment at FHML. Her interests are in competency-based education and assessment, and expertise development in professional education. She is involved in design and implementation of e-portfolios and work-based assessment programmes in (undergraduate and postgraduate) health professions education (medical training in particular). Her research focuses on various aspects of competency-based education and assessment, and more specifically on assessor cognition, work-based assessment and programmatic assessment in education for the (health) professions. She received several awards for her work as a researcher (AERA, Netherlands Association for Medical Education) and medical educator.

Jeroen (H.H.L.M.) Donkers was born in Gemert in The Netherlands in 1963. He received M.Sc. in knowledge engineering (artificial intelligence and operations research) from Maastricht University in 1997 and Ph.D in artificial intelligence in 2003 at the same university. From 1987 till 2004 he was scientific programmer and teacher at Maastricht University, and since 2004 assistant professor, first at the computer science department and since 2007 at SHE.

The topic of his PhD thesis was on opponent modelling in computer game-playing. At the department of Computer Science he has been working on reasoning with uncertainty in several domains. He introduced the EPASS portfolio system in 2009 to Maastricht University (www.epass.eu). He has been project leader of a nationally (SURF) funded project on improving the ICT support of progress testing (VGTogether). He has been involved in the EU-funded WATCHME project with a main focus on student modelling and learning analytics (www.project-watchme.eu). He has supervised MSc and PhD students in computer science and in educational research on several topics.

Currently Dr. Jeroen Donkers acts as psychometrics consultant for (digital) assessment as well as for educational research in SHE.

Peter van Rosmalen received his M.Sc. degree in Mathematics from the University of Leiden in 1981 and his PhD in Technology Enhanced Learning in 2008 from the Open University of the Netherlands.

He has worked in both business and universities. Currently, he is associate professor and chair of the taskforce 'Instructional Design and E-learning' at the Department of Educational Development and Research of the Faculty of Health, Medicine and Life Sciences at Maastricht University. His research focuses on how to use technology to empower learner and teacher within topics such as simulations and serious games, e-universities, MOOCs, computer supported cooperative learning, adaptive e-learning, peer support, learning networks, creativity, sensors in education, and language technologies for learning.