Advances in Medical Education and Practice

Dovepress

open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

# Leveraging Narrative Feedback in Programmatic Assessment: The Potential of Automated Text Analysis to Support Coaching and Decision-Making in Programmatic Assessment

Balakrishnan R Nair[1], Joyce MW Moonen - van Loon[2], Marion van Lierop[3], Marjan Govaerts[2]

[1]University of Newcastle, Centre for Medical Professional Development, Newcastle, Australia; [2]School of Health Professions Education, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, the Netherlands; [3]Department of Family Medicine, Faculty of Health Medicine and Life Sciences, Maastricht University, Maastricht, the Netherlands

Correspondence: Joyce MW Moonen - van Loon; Balakrishnan R Nair, Email j.moonen@maastrichtuniversity.nl; kichu.nair@newcastle.edu.au



**Introduction:** Current assessment approaches increasingly use narratives to support learning, coaching and high-stakes decision-making. Interpretation of narratives, however, can be challenging and time-consuming, potentially resulting in suboptimal or inadequate use of assessment data. Support for learners, coaches as well as decision-makers in the use and interpretation of these narratives therefore seems essential.

**Methods:** We explored the utility of automated text analysis techniques to support interpretation of narrative assessment data, collected across 926 clinical assessments of 80 trainees, in an International Medical Graduates' licensing program in Australia. We employed topic modelling and sentiment analysis techniques to automatically identify predominant feedback themes as well as the sentiment polarity of feedback messages. We furthermore sought to examine the associations between feedback polarity, numerical performance scores, and overall judgments about task performance.

**Results:** Topic modelling yielded three distinctive feedback themes: Medical Skills, Knowledge, and Communication & Professionalism. The volume of feedback varied across topics and clinical settings, but assessors used more words when providing feedback to trainees who did not meet competence standards. Although sentiment polarity and performance scores did not seem to correlate at the level of single assessments, findings showed a strong positive correlation between the average performance scores and the average algorithmically assigned sentiment polarity.

**Discussion:** This study shows that use of automated text analysis techniques can pave the way for a more efficient, structured, and meaningful learning, coaching, and assessment experience for learners, coaches and decision-makers alike. When used appropriately, these techniques may facilitate more meaningful and in-depth conversations about assessment data, by supporting stakeholders in interpretation of large amounts of feedback. Future research is vital to fully unlock the potential of automated text analysis, to support meaningful integration into educational practices.

**Keywords:** programmatic assessment, narrative feedback, learning analytics, text mining, international medical graduates

## Introduction

Current conceptualizations of assessment emphasize the dual purpose of robust decision making about learners' progress as well as guiding learners' competence development. In recent decades, the importance of timely and constructive feedback in learning and assessment has been underscored, with the adage "no assessment without feedback" emerging as a fundamental principle.

Current assessment approaches within programmatic assessment (PA), emphasize a learner-centered approach and assessment for learning. Multiple low-stakes assessments generate comprehensive feedback, supporting longitudinal coaching and the development of relevant competencies.[1] Diverse assessment tools used by various



Received: 21 February 2024
Accepted: 27 June 2024
Published: 15 July 2024

assessors across multiple contexts generate assessment data that provide a holistic understanding of a learner's progress and achievements. Increasingly, assessments are based on narratives (qualitative data), offering more meaningful feedback than grades or scores (qualitative assessment data) and facilitating productive coaching conversations. This enhances the credibility and transparency of decision-making processes[2] and supports learners in becoming self-directed professionals.[3,4] However, the interpretation of narratives can be difficult due to conflicting messages, hedging, the volume of feedback, and differences in language used by feedback providers.[5–7] Interpretation of feedback becomes particularly challenging when narrative data are aggregated over time and across multiple assessment occasions and assessors, as in programmatic or competency-based assessment programs.

Efforts have been made to aid individual learners, coaches, and decision-makers in interpreting narrative feedback through various means, such as combining or triangulating with quantitative data and use of scoring rubrics or automated text analysis.[8–11] Automated text analysis holds promise as a valuable tool to support feedback processes. An initial exploratory technical study[12] demonstrated the technical feasibility of harnessing "topic modelling" and "sentiment analysis" to generate comprehensive overviews of the primary topics embedded within narrative feedback data. Study findings furthermore showed that this approach enabled the evaluation of the sentiment polarity (positive, neutral, or negative), associated with these identified topics. Notably, the study was conducted in a large scale undergraduate medical curriculum using programmatic assessment approaches and encompassed all the narrative feedback uploaded to students' portfolios (N = 1516), thus providing a robust foundation for further research and exploration. More specifically, there is still much to learn about how text analytics and computer-based analysis can genuinely enhance the assessment experience of our learners, by providing clear (just-in-time) and accurate overviews of their specific strengths and weaknesses. Additionally, further research on relationships between qualitative and quantitative assessment data, ie if and how narrative feedback aligns with performance scores, may enhance our understanding of how to use and interpret narratives in current assessment approaches.

The primary objective of this research is to investigate the characteristics of narrative assessment data within International Medical Graduates' (IMGs) portfolios by employing automated text analysis techniques. Specifically, we focus on discerning predominant feedback topics and sentiment polarity within these topics across various clinical disciplines (Adult Medicine, Adult Surgery, Women's Health, Child Health, Mental Health, and Emergency Medicine). The aim is to uncover potential variations in feedback patterns across distinct topics and/or disciplines. Additionally, we seek to examine the associations between feedback sentiment polarity, numerical scores, and overall judgments about task performance.

As this research is intended to uncover the potential of automated text analysis, we furthermore present a suggestion of how individual learners and other stakeholders could be supported by employing these text analysis techniques, facilitating a more profound understanding of the feedback provided.

# Methods

## Setting

The context for our study is one of the longitudinal assessment programs for the licensing requirements of International Medical Graduates (IMGs) in Australia. The IMGs originate from many different countries including Egypt, India, Sweden, South Africa, Brazil, Sri Lanka, Pakistan, Myanmar Burma, Iraq, among others. The assessment program uses Workplace Based Assessments (WBAs) to evaluate IMGs, utilizing diverse assessment methods including Mini-Clinical Examinations (Mini-CEXs), Case-Based Discussions (CBDs), and Multisource Feedback (MSF). Over the past decade, more than 350 IMGs participated in the assessment. The assessment program has proven cost-effective and has received positive feedback from learners and assessors.[13,14]

Assessors of Mini-CEXs and CBDs attended a two-hour compulsory calibration session to ensure the feedback was appropriate and aligned with the numerical scores. The assessment forms utilized in this study can be found in the Appendix.

## Data Collection

The dataset consists of all assessments submitted between November 2020 and October 2022 after an assessor's direct observation of the IMG's performance, either in clinical skills (Mini-CEX) or during a structured discussion of a patient problem (CBD). The Multisource Feedback was omitted from this study because of the anonymous identity and various professions of respondents (who are not all calibrated assessors in the program), coupled with the protracted duration over which this feedback is dispensed rather than confined to a single occurrence.

The Mini-CEX and CBD assessments represent IMG performance feedback in six different disciplines: Adult Medicine, Adult Surgery, Women's Health, Child Health, Mental Health, and Emergency Medicine. Each form required assessors to evaluate trainee performance using a predefined set of criteria and a 5-point rating scale organized in three levels of "below expectations" (1–2), "meets expectations" (3) and "exceeds expectations" (4–5) at the standard of an Australian graduate at the end of the first postgraduate year (PGY1). Additionally, assessors were asked to provide an overall global rating of the candidate's performance and professionalism in all areas, distinguishing between "Competent" and "Not Competent". The narrative feedback provided by the assessor was intended to complement the numerical scores, offering a comprehensive assessment of the candidate's performance.

Before data analysis, the dataset underwent cleaning, removing HTML tags, trimming white spaces, and replacing special characters from the narrative feedback. Since feedback comments often comprised multiple sentences, each possibly addressing different aspects, all feedback comments were segmented into individual sentences.

## Text Analysis: Topic Modelling

Topic modelling is a statistical and computational technique used in natural language processing and machine learning to identify underlying topics or themes in a collection of texts. It operates on the assumption that each document in the collection is a mixture of different topics, and each word is attributable to one of those topics. We used Latent Dirichlet Allocation (LDA)[15,16] to identify the words that belong to a specific topic and the mixture of topics that described each document simultaneously. This proven text mining technique is commonly applied in analysis of, for example, a large amount of news articles, or for characterizing a scientific field's research.[17–20] Utilizing a topic modelling software package,[21] we meticulously examined the topics automatically identified in the dataset, which resulted in identification of three distinctive main topics: Medical Skills, Knowledge and Communication & Professionalism. These topics were then used in a semi-supervised topic modelling, in which the model utilizes a small amount of labelled data (documents with predefined topics or themes), called seeds, alongside a larger pool of unlabelled data. The labelled data help guide the learning process of the topic modelling algorithm, while the unlabelled data allow for the discovery of latent topics. In other words, semi-supervised topic modelling assigns a higher probability to sentences containing seed words, ensuring they are more likely to be categorized under a specific topic.

For this study, using the results of the unsupervised topic modelling and word (co)occurrences in the feedback, we defined the following seeds as input for the semi-supervised topic modelling:

- Medical Skills: management, manage, plan, history, hx, physical, examination, skill, symptom, sx, treatment, tx, clinical, documentation, document
- Knowledge: judgement, diagnosis, understanding, knowledge, assessment, investigation, differential, differential diagnosis, ddx, differentials, knowledgeable, dx, rationalise, deduce, induce
- Communication & Professionalism: communication, communicator, communicate, counselling, counsel, interview, interviewing, rapport, approach, question, conduct, interaction, humanistic, professional, professionalism, social, discuss, empathy, empathic, empathetic, engage, explain, clarify, explanation, sentence, organised, English, record, legible, approachable, manner, mannered

The semi-supervised topic modelling yielded a partition of the narrative feedback sentences over the topics. We investigated the distribution of sentences among the topics and across the different disciplines.

## Text Analysis: Sentiment Analysis

Sentiment analysis is the computational study of opinions, sentiments and emotions expressed in text.[22] Sentiment analysis is 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.[18,22–25] In this study, the sentiment analysis tool Grasp was applied to each sentence, automatically assigning a polarity ranging from +1 (indicating a positive sentiment) to −1 (indicating a negative sentiment), or 0 for neutral sentiments, based on the words within the sentence and their co-occurrences. Grasp is a Python tool that was selected based on its performance and possibilities for including context-specific deficiencies in the language model, eg "skillful" and "expected level", alongside the predetermined dictionary.

We calculated the average sentiment polarities of sentences per topic and across the different disciplines. We furthermore calculated the average sentiment polarity and the standard deviation for all feedback sentences in assessments graded as "Competent" and assessments graded as "Not Competent" respectively. Using a *t*-test, we tested whether the averages of both sets significantly differed, thereby determining whether these feedback sentences came from two separate sets, which would suggest that the average sentiment polarity of the narratives signifies competent performance (or not), even without an explicit overall global rating.

## Relationship Between Performance Scores and Narratives

Examining the associations between feedback sentiment polarity, numerical scores, and overall judgments about task performance, requires a comparison between numerical scores given on a 1-to-5-scale to the sentiment polarity ranging from −1 to 1 on the same set of assessments. To compare two sets of values (in this case, numerical performance scores and sentiment polarities), the values should be corresponding to the same scale. Therefore, we calculated the *polarity grade* of each assessment as the grade that would be given based on the sentiment polarity of the narrative feedback on a 1-to-5-scale. A direct comparison between sentiment polarities (ranging from −1 to 1) and performance scores (ranging from 1 to 5) was facilitated by a linear transformation, where the polarity grade was computed as 2*(sentiment polarity) + 3. Thus, a sentiment polarity of −1 corresponds to a performance score of 2*-1+3=1, a sentiment polarity of 0 corresponds to score 2*0+3=3 and a sentiment polarity of +1 to score 2*1+3=5.

We then calculated correlations between the resultant polarity grade and the numerical score. The same analysis was repeated for feedback sentences collected on assessments that are graded "Competent" and "Not Competent" respectively.

Assessors numerically graded the performance on several assessment criteria (see Appendix). Therefore, we connected each assessment criterion to the topics identified from the complete dataset, resulting from the topic modelling. We then determined the average performance score per topic and compared this to the average polarity grade of the feedback on that same assessment assigned to the same topic. We executed the same steps as above for each identified topic.

# Results

The dataset comprises 682 mini-CEXs and 353 CBDs, totaling 1035 clinical assessments. Assessments without accompanying comments were omitted from the analysis, resulting in a dataset of 603 Mini-CEXs and 323 CBDs from 80 candidates. Table 1 presents the descriptive details of the assessments in the dataset. Of the assessments, 211 were in Adult Medicine, 163 in Adult Surgery, 130 in Women's Health, 110 in Child Health, 129 in Mental Health and 183 in Emergency Medicine. In 891 assessments, the assessor's overall judgement was "Competent", whereas the assessor's overall judgement was "Not Competent" in 35 forms: 7 CBDs and 28 mini-CEXs. The average number of words in the 891 assessments with an overall judgement "Competent" was 26.51 and in the 35 assessments graded "Not Competent" 67.26. For the disciplines of Mental Health, Child Health and Women's Health, there are much fewer CBDs than Mini-CEXs.

## Topic Modelling

The feedback on all 926 assessments was written in 2533 sentences. Each sentence was automatically categorized into one of the three topics. Some sentences could not be assigned to a topic (eg "no obvious deficiencies"), in which case the feedback was labelled as "General".

**Table 1** Overview of the Number of Assessments in Total and per Type and the Average Number of Words Used in the Feedback

| | Total | | | | Competent | | | | Not Competent | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Avg #Words** | **# Assessments** | **#Mini-CEX** | **#CBD** | **Avg #Words** | **# Assessments** | **#Mini-CEX** | **#CBD** | **Avg #Words** | **# Assessments** | **#Mini-CEX** | **#CBD** |
| **Adult Medicine** | 27.04 | 211 | 95 | 116 | 25.91 | 203 | 88 | 115 | 55.75 | 8 | 7 | 1 |
| **Emergency Medicine** | 32.83 | 183 | 108 | 75 | 30.18 | 179 | 106 | 73 | 151.50 | 4 | 2 | 2 |
| **Mental Health** | 32.71 | 129 | 94 | 35 | 29.43 | 117 | 83 | 34 | 64.67 | 12 | 11 | 1 |
| **Child Health** | 24.05 | 110 | 98 | 12 | 24.07 | 109 | 97 | 12 | 22.00 | 1 | 1 | 0 |
| **Women's Health** | 25.71 | 130 | 103 | 27 | 24.66 | 124 | 97 | 27 | 48.67 | 6 | 6 | 0 |
| **Adult Surgery** | 24.80 | 163 | 105 | 58 | 24.09 | 159 | 104 | 55 | 53.00 | 4 | 1 | 3 |
| **Total** | 28.05 | 926 | 603 | 323 | 26.51 | 891 | 575 | 316 | 67.26 | 35 | 28 | 7 |



https://doi.org/10.2147/AMEP.S465259 DovePress

Table 2 presents the distribution of feedback sentences, across clinical disciplines and the various topics as identified by the semi-supervised topic modelling. As shown, 47.69% of all sentences were automatically assigned to Medical Skills, 28.07% to Knowledge and 23.10% to Communication & Professionalism. Most feedback sentences were provided on assessments in Adult and Emergency Medicine (both ~21%) and least in Child Health.

For the topic Communication & Professionalism, the number of sentences from CBDs (71) is a lot lower than the number of sentences from Mini-CEXs (514).

## Sentiment Analysis

Table 3 presents the average polarities of the assessments based on the automatic assignment of the polarity (between −1 and +1) to the sentences within these assessments.

When observing the results per type of assessment, we see that on the CBDs, the average polarity for Medical Skills (0.3052) and Knowledge (0.2841) is higher than for the Mini-CEXs (0.2486 and 0.2106, respectively). In contrast, for Communication & Professionalism, it is the other way around (0.2757 for Mini-CEXs and 0.1968 for CBD).

Table 4 presents the average polarities of the assessments graded "Competent" based on the automatic assignment of the polarity (between −1 and +1) to the sentences within these assessments. The average polarity of the 2366 sentences from assessments graded "Competent" is 0.2795, with a standard deviation of 0.3130. 68% of the polarities lie within one standard deviation of the mean and only 4% lie outside two standard deviations of the mean.

The same results for the assessments graded "Not Competent" are presented in Table 5. The average polarity of the 167 sentences from assessments graded "Not Competent" is 0.0050, with a standard deviation of 0.2652. 75% of the polarities lie within one standard deviation of the mean and 5% lie outside two standard deviations of the mean.

**Table 2** Distribution of Feedback Sentences to Topics and Disciplines

| DISCIPLINE – TOPIC | Medical Skills | Knowledge | COM. & Prof. | General | all |
|---|---|---|---|---|---|
| **Adult Medicine** | 11.61% | 6.59% | 2.96% | 0.16% | 21.32% |
| **Emergency Medicine** | 9.63% | 6.99% | 4.11% | 0.28% | 21.00% |
| **Mental Health** | 6.59% | 3.08% | 5.01% | 0.20% | 14.88% |
| **Child Health** | 4.18% | 1.46% | 4.18% | 0.12% | 9.95% |
| **Women's Health** | 5.61% | 5.13% | 3.63% | 0.12% | 14.49% |
| **Adult Surgery** | 10.07% | 4.82% | 3.20% | 0.28% | 18.36% |
| **all** | 47.69% | 28.07% | 23.10% | 1.14% | 100.00% |

**Table 3** The Average Polarity of the Sentences in the Disciplines Assigned to the Various Topics

| Discipline – Topic | Medical Skills | Knowledge | COM. & Prof. | General | all |
|---|---|---|---|---|---|
| **Adult Medicine** | 0.2602 | 0.2281 | 0.1931 | 0.1167 | 0.2399 |
| **Emergency Medicine** | 0.2966 | 0.2900 | 0.3265 | 0.1905 | 0.2989 |
| **Mental Health** | 0.1938 | 0.2296 | 0.2185 | 0.1933 | 0.2095 |
| **Child Health** | 0.2233 | 0.2404 | 0.2618 | 0.1667 | 0.2413 |
| **Women's Health** | 0.3309 | 0.1916 | 0.3184 | 0.3333 | 0.2784 |
| **Adult Surgery** | 0.2952 | 0.2695 | 0.2775 | 0.1500 | 0.2831 |
| **All** | 0.2708 | 0.2447 | 0.2661 | 0.1833 | 0.2614 |

**Table 4** Average Polarity of the Sentences on Assessments Graded "Competent" Provided in the Different Disciplines Assigned to the Various Topics

| Discipline – Topic | Medical Skills | Knowledge | COM. & Prof. | General | all |
|---|---|---|---|---|---|
| **Adult Medicine** | 0.2667 | 0.2506 | 0.1949 | 0.1167 | 0.2507 |
| **Emergency Medicine** | 0.3241 | 0.3178 | 0.3372 | 0.2222 | 0.3234 |
| **Mental Health** | 0.2255 | 0.2561 | 0.2434 | 0.1933 | 0.2375 |
| **Child Health** | 0.2328 | 0.2404 | 0.2643 | 0.1667 | 0.2464 |
| **Women's Health** | 0.3406 | 0.2134 | 0.3246 | 0.3333 | 0.2931 |
| **Adult Surgery** | 0.3049 | 0.3089 | 0.3051 | 0.1917 | 0.3044 |
| **All** | 0.2869 | 0.2702 | 0.2792 | 0.2006 | 0.2795 |

**Table 5** Average Polarity of the Sentences on Assessments Graded "Not Competent" Provided in the Different Disciplines Assigned to the Various Topics

| Discipline – Topic | Medical Skills | Knowledge | COM. & Prof. | General | all |
|---|---|---|---|---|---|
| **Adult medicine** | 0.1073 | −0.2198 | 0.1625 | | 0.0075 |
| **Emergency medicine** | −0.0481 | 0.0100 | 0.1786 | 0.0000 | 0.0129 |
| **Mental health** | −0.0395 | −0.0889 | 0.0848 | | 0.0081 |
| **Child health** | −0.1000 | | 0.0000 | | −0.0750 |
| **Women's health** | −0.1167 | −0.0233 | 0.2429 | | 0.0486 |
| **Adult surgery** | 0.0296 | −0.0917 | −0.0143 | −0.1000 | −0.0356 |
| **All** | −0.0116 | −0.0650 | 0.1130 | −0.0500 | 0.0050 |

The average sentiment for all sentences in "Competent" assessments is 0.2795, and in "Not Competent" assessments, it is 0.0050, with the largest difference of 0.3352 in the topic "Knowledge". When looking at the series of sentences from the "Competent" and "Not Competent"-grades assessments, we get a t-statistic of 11.06, leading to $P < 0.0001$, concluding that the means are significantly different. Further evaluations of the sentiment-polarity per topic yield t-statistics of 7.5, 7.7 and 3.6 for Medical Skills, Knowledge and Communication & Professionalism, indicating that the means per topic significantly differ between the assessments with overall global ratings "Competent" and "Not Competent".

## Narrative feedback's sentiment polarity vs Numerical performance scores

For this dataset and algorithm, the linear transformation from polarities to a 1-to-5 scale yielded an average absolute deviation between polarity grade and performance score of 0.52, with a standard deviation of 0.37 and a maximum deviation of 2. An example of this maximum deviation is an assessment with an average grade of 3 and narrative feedback "very good record, good history", receiving a sentiment polarity of +1, translating to a polarity grade of 5.

Various other linear and polynomial transformations (Table 6), even when based on the trend of data points and knowledge of the dataset (eg the minimum average assessment score in the dataset is 1.5 and the minimum polarity is −0.4, and the maxima are 5 and +1 respectively), gave no improvements on the prediction of the assessment score based on the sentiment polarity.

**Table 6** Average, Standard Deviation and Maximum of the Absolute Difference Between the Average Performance Score and the Calculated Polarity Grade for Various Polynomial Transformations

| Transformation of Sentiment Polarity *p* into "Polarity Grade" *x* | Based on the Absolute Difference Between the Average Performance Score to the Calculated "Polarity Grade" *x*. | | |
|---|---|---|---|
| | Average | Std. dev. | Max. |
| $x=2p+3x$ | 0.5209 | 0.3737 | 2 |
| $x=2.5p+2.5$ | 0.5615 | 0.4357 | 2.4167 |
| $x=12.121p_2-0.0866p+1.9979$ | 0.9464 | 0.5985 | 2.7424 |
| $x=-4.908p^3+3.616p^2+3.335p+2.25$ | 0.6755 | 0.4831 | 2.6827 |
| $x=3.4428p+2.4157$ | 0.6427 | 0.5011 | 2.6581 |

Next, we plotted the average sentiment polarity of assessments with a similar (rounded) average performance score in Figure 1 (blue line). The upward trend is clearly visible. The correlation between the average performance grades and the average polarities is 0.9360.

The same plot looking at the data exclusively for the assessments that have an overall global rating "Competent" is presented by the orange line in Figure 1, and the plot for the assessments that with overall global rating "Not Competent" is shown in grey. We found an upward trend for the "Competent" assessments, with the lowest (rounded) average score of 2, with a correlation of 0.9803. This result does not hold for the "Not Competent" assessments, with the highest (rounded) score of 3, with a correlation of 0.5527. However, it should be noted that only 3.78% of the assessments are graded "Not Competent", so each outlier greatly affects the overall output.

The assessment criteria on which numerical grades were provided on the Mini-CEXs and CBDs, as presented in the Appendix, can be categorized into the same topics found in the narrative feedback. For "Medical Skills", criteria "History-taking skills" and "Physical examination skills" were used for Mini-CEXs and criterion "Management plan" for the CBDs. For "Knowledge", criterion "Clinical judgement / clinical reasoning" was used for Mini-CEXs and criteria "Differential diagnosis and summary list" and "Clinical judgement / clinical reasoning" for CBDs. For "Communication & Professionalism", criteria "Medical interviewing / Communication skills" and "Professionalism / humanistic skills" were used for Mini-CEXs and criterion "Clinical record keeping" for CBDs.

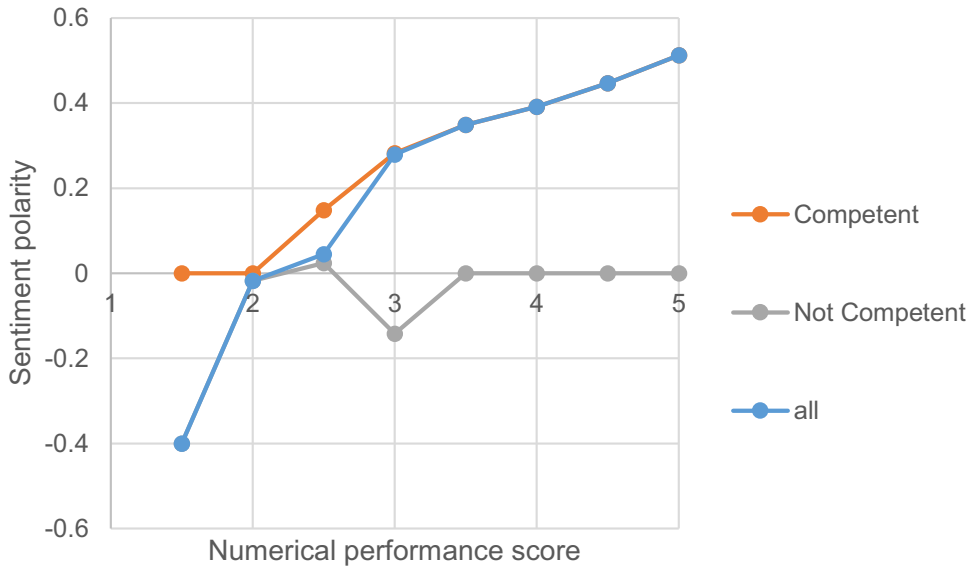


**Figure 1** The average sentiment polarity (y-axis) of all assessments with the same (rounded) average numerical score (x-axis).

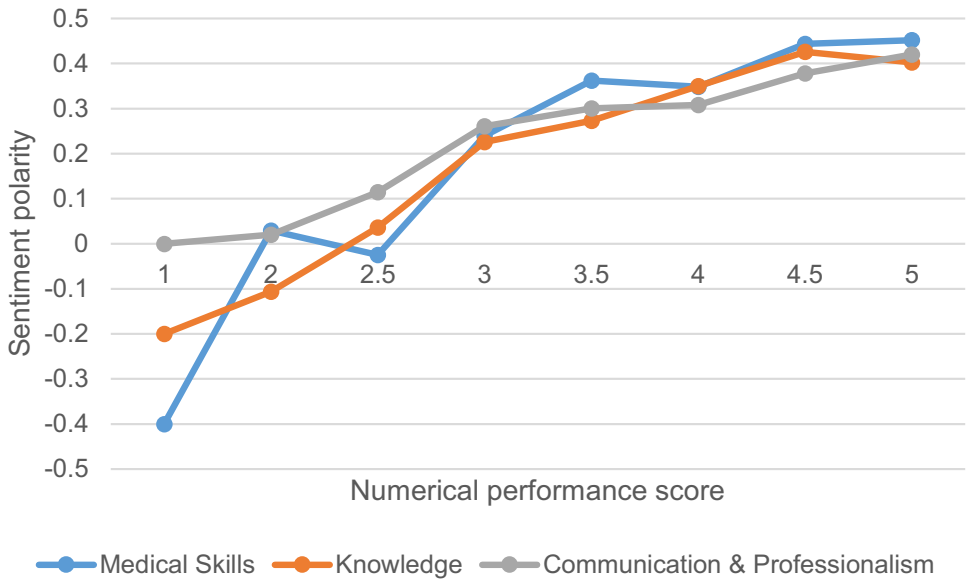


**Figure 2** All assessments' average sentiment polarity (y-axis) with the same rounded average numerical score (x-axis) per topic.

Recreating the graph from Figure 1, the results per topic are presented in Figure 2.

As an example for clarification, on one particular Mini-CEX assessment, seven feedback sentences are written; two of them are automatically assigned to topic "Medical Skills" with an average sentiment polarity of 0.35, four to "Knowledge" with an average sentiment polarity of 0.125 and one to "Communication & Professionalism" with sentiment polarity −0.1. The average performance score on criteria "History-taking skills" and "Physical examination skills" on that Mini-CEX is equal to 4, criterion "Clinical judgement / clinical reasoning" is scored with a 3, and the average performance score on criteria "Medical interviewing / Communication skills" and "Professionalism / humanistic skills" is equal to 4. This one assessment adds to the average data of each of the three series presented in Figure 2.

The correlation between the grades (x-axis) and the average polarities (y-axis) is 0.9466 for "Medical Skills", 0.9691 for "Knowledge" and 0.9594 for "Communication & Professionalism".

## Discussion

This study aimed to investigate the characteristics of assessment data within International Medical Graduates' (IMGs) portfolios by employing automated text analysis techniques to identify predominant feedback topics and sentiment polarity within these topics across various clinical disciplines.

In general, we saw that assessors used more words to motivate or instruct the candidate if they felt the candidate did not reach competence. In case of incompetent performance, assessors may have felt the need to substantiate their overall judgement (for reasons of accountability) and/or to provide trainees with meaningful feedback for further learning. The nature of these narratives, including constructive feedback, likely requires more detailed and specific information. When focusing on the topics in the narrative feedback, we found that most feedback is provided on "Medical Skills", especially in the disciplines of Adult Medicine, Emergency Medicine and Adult Surgery. In the disciplines of Mental Health and Child Health, however, feedback on "Communication & Professionalism" is quantitatively comparable to feedback on "Medical Skills". In these disciplines, heightened attention to communication and professionalism arises from the imperative for patient-centered care, interdisciplinary collaboration, the vulnerability of patients, the necessity for informed decision-making, and the profound impact on treatment outcomes, reflecting the unique challenges and ethical considerations inherent in these fields. This emphasis underscores the pivotal role of interpersonal skills alongside clinical expertise in ensuring holistic and effective healthcare provision. For Women's Health, on the other hand, feedback on "Knowledge" is documented more frequently. This was not unexpected since the overall impression and feedback from the assessors is that most IMGs coming to Australia are not adequately trained in women's health. This

may be related to cultural issues in their home countries. Understanding these feedback variations can guide future assessment strategies and highlight areas where more attention to providing feedback may be needed, or where setting up training programs in specific disciplines is necessary to support trainees' competence development.

We found that the assessment criteria on the Mini-CEX and CBD, in general, reflect the topics that were automatically derived from the complete data set. However, for some sentences automatically determining the topic is arbitrary, for example with the feedback sentence "Able to discuss management of various issues in a very practical way". Also, automatically determining sentiment polarity can be challenging. For example, the feedback on an assessment, with overall global rating "Not competent", contains two sentences:

- "Good patient interaction."
- "Likely to be a good doctor but is at present entirely unfamiliar with antenatal care, diabetes, vaccine safety, etc."

The first sentence gets a sentiment score of 0.7, the second 0.075. The second sentence includes a very strong concern ("entirely unfamiliar") but is preceded by positive "likely to be a good doctor", which "cancels out" the negativity in the automatically assigned sentiment polarity. Averaging over both sentences, the sentiment polarity of the narrative feedback is 0.3875, which is quite positive. "Not competent" assessments may thus receive a quite positive average sentiment polarity in their narrative feedback, when assessors write a very positive comment about the candidate's performance, before (sometimes) carefully describing the areas of improvement. This finding underlines the need to acknowledge the human dimensions in feedback processes, both in provision of feedback (and thus the nature of narrative assessment data, such as use of veiled language) as well as "meaning giving" to feedback (eg contextualization of specific comments). The algorithms underlying automated text analysis may help signify key strengths, weaknesses, and omissions in feedback, but do not generate "the absolute truth"; human interpretation and processing of feedback remain necessary to optimize the credibility and utility of feedback.

Based on our findings, suggestions for improving the feedback practices in the IMG program include modification of assessment instruments to request assessors to identify strengths and areas for improvement and concerns more explicitly and separately, as this could enhance the richness and clarity of feedback, and improve the performance of the automated text mining analyses. Furthermore, introducing leading suggestions on the assessment forms to cover specific topics such as Medical Skills, Knowledge, and Communication & Professionalism could provide a structured framework for assessors, ensuring a more comprehensive and balanced assessment of the main topics in education. Finally, training assessors to write complete English sentences was identified as a crucial area for improvement. For example, the feedback "good content and flow" was not deductible automatically because it is unclear what good content and flow means, as this could refer to eg communication, the medical procedure, investigation or history taking, or a written or verbal presentation. Clear, complete sentences enable better comprehension of feedback and ensure that the feedback is actionable and meaningful.

The study explored the relationship between numerical scores and sentiment polarity assigned through automated text analysis. For an individual assessment, we found no straightforward relationship between the numerical grades and the algorithmically assigned sentiment polarity. These findings are in line with previous research showing that there is not always a 1-to-1 correspondence between performance scores and documented narrative feedback. Performance scores and narratives therefore seem to provide supplementary information about a trainee's performance.[2,26–28] Our findings however show a strong positive correlation between the average numerical scores and the average algorithmically assigned sentiment polarity, suggesting that, in general, higher performance scores are coupled with more positive narrative feedback and vice versa. These findings illustrate the potential of automated text analysis strategies in extracting meaningful insights from aggregated narrative feedback, providing a valuable tool to support interpretation of diverse and subjective performance evaluations. Consequently, automated text analysis emerges as a valuable and complementary method for gaining deeper insights into the evaluative process and contributing to a more holistic and informative assessment.

Findings from our study, however, also reflect that while automated language models are smart, their results are based on probabilistic estimation and are therefore never completely accurate. When applied carefully and appropriately, ie

acknowledging the importance of human interpretation and human dimensions of feedback, application of these models can, however, provide direction to reflection, learning, coaching as well as decision making processes.

There are several limitations in this study. First, the exclusion of multisource feedback data limited the comprehensiveness of the analysis. Future research should aim to incorporate or compare the text analyses of other assessments to this data for a more holistic view. Second, topic modelling's dependency on co-occurrences of words in sentences posed a challenge when assigning sentences to topics. Sentences are assigned to topics with a certain probability. When using topic modelling to interpret or assess the feedback for an individual learner, it is important to realize that the output might not be entirely correct and personal interpretation of the output remains necessary. Third, the definition of the sentiment polarity was open to interpretation in some instances. For example, the feedback sentences "knew the essentials" and "no obvious deficiencies" are both assigned a neutral polarity. The texts do not clearly give a direction: is knowing the essentials the bare minimum, or is it a positive sign of performance? If there are no obvious deficiencies, are there any other deficiencies, and does that make the performance satisfactory or not? This highlights the complexity of sentiment analysis and the need for further refinement, standardization, and adjustment of the assessment instruments.

With the availability of the presented text analyses in this paper, we present a suggestion to support an individual learner to use their feedback in the learning process by showing strengths, weaknesses and omissions in the narrative feedback as compared to the underlying educational framework, helping students and coaches in developing reflections and new learning objectives. As an example, for the current setting, an overview as presented in Figure 3 is a feasible possibility, showing for one IMG the sentences per topic and for each sentence some metadata on the assessment it came from (eg the date of the assessment and the complete feedback text to present context if needed). Then, as the topic modelling assigns a probability of the match between the sentence and the topic, this fit-percentage

| Sentiment Polarity | Feedback | Fit | Score | Discipline | Global Rating |
|---|---|---|---|---|---|
| **Medical Skills** | | | | | |
| | the ability to form a plan requires some aspect of a competent history and examination | 100% | 2.5714 | Emergency Medicine | Not Competent |
| | fortunately the management plan was safe as he is deferring to the inpatient services, but the ability to communicate this patients issues may lead to significant gaps when being handed over | 75% | 2.5714 | Emergency Medicine | Not Competent |
| | good clinical judgement demonstrated when the patient became unwell with excellent initial management and resuscitation | 73% | 3.0000 | Adult Surgery | Competent |
| | we discussed the virtues of a direct clear assessment and clear management plan | 72% | 3.0000 | Adult Medicine | Competent |
| | good overall history | 48% | 3.0000 | Child Health | Competent |
| | the essential history and exam points we discussed were not followed during this wba | 39% | 2.5714 | Emergency Medicine | Not Competent |
| | good history taking skills while validating her feelings | 38% | 4.8333 | Mental Health | Competent |
| | these skills are below those i expect of a first term intern | 32% | 2.5714 | Emergency Medicine | Not Competent |
| **Knowledge** | | | | | |
| | very good presentation of history and differential diagnosis | 99% | 3.0000 | Adult Surgery | Competent |
| | not so clear on the different imaging modalities best suited for obtaining the differential diagnosis but was able to navigate with prompting | 94% | 3.0000 | Adult Surgery | Competent |
| | overall approach to assessment & diagnosis was reasonable and this included appropriate discussion with different specialties | 73% | 3.0000 | Mental Health | Competent |
| | comprehensive record keeping with detailed notes, well conducted examination and well thought differential in an unwell complex patient, advised to read more on antibiotics choices for sepsis of unknown origin to further enhance knowledge of microbiology | 65% | 3.2500 | Adult Medicine | Competent |
| | appropriate plan for admission and further investigation with further sub specialty review | 50% | 2.5714 | Emergency Medicine | Not Competent |
| | was able with prompting to differentiate between biliary colic, cholecystitis and cholangitis and their differential management | 47% | 3.0000 | Adult Surgery | Competent |
| | able to express few good differential diagnoses with minimal prompting | 42% | 3.1667 | Emergency Medicine | Competent |
| | the investigations were appropriate | 36% | 2.5714 | Emergency Medicine | Not Competent |
| | good knowledge of psychiatry | 26% | 4.8333 | Mental Health | Competent |
| **Communication & Professionalism** | | | | | |
| | history taking appropriate maintained professionalism counselled appropriately | 42% | 3.0000 | Women's Health | Competent |
| | engaged patient well to establish good rapport | 42% | 4.8333 | Mental Health | Competent |
| | medical interviewing needs to be improved | 36% | 2.1667 | Adult Surgery | Not Competent |
| | established good rapport with patient | 35% | 2.5714 | Emergency Medicine | Not Competent |
| | approachable | 33% | 3.1667 | Emergency Medicine | Competent |
| **General** | | | | | |
| | averagely competent | 0% | 3.1667 | Emergency Medicine | Competent |
| | weakness, urinary retention and age | 0% | 2.5714 | Emergency Medicine | Not Competent |

**Figure 3** Suggestion of a presentation of structured narrative feedback to a candidate or assessor, ordered in the established topics. The polarity is presented on a horizontal bar, the feedback sentence, how well the sentence fits the topic, the average score (1–5) of the assessment to which this feedback belongs, the discipline of the assessment, and whether the candidate was assessed "Competent" or "Not Competent".

can be displayed, the sentiment polarity (between −1 and +1) presented in a small bar chart, the average score of the numeric (1–5) items on the assessment, the discipline of the assessment, and the overall global rating. Such an overview can give insight into the number of assessments per topic (ie the number of sentences) and help identify personal strengths and weaknesses by observing the polarities. However, further research needs to be done to evaluate actual usability and support of the interpretation or assessment of such an overview for the individual learner and assessor.

In conclusion, our study aligns with recent debates in education and assessment regarding the use, opportunities, and risks of artificial intelligence (AI).[29] With the recent rise of Large Language Models (LLMs), the developments of language processing techniques give more and more opportunities for improving the accuracy of automated text analyses. Our findings highlight the transformative role that automated text analysis techniques can play in the education's assessment and feedback landscape. By embracing these technologies yet addressing their nuances, we can pave the way for a more efficient, structured, and meaningful educational experience for learners and decision-makers alike. Future research and advancements in this domain are vital to fully unlock the potential of automated text analysis and to ensure responsible integration into educational practices.

## Ethics Statement

All candidates and assessors gave consent to use their data for research purposes. The research was approved by the Health Services Research and Ethics committee of the Health Service (approval number A.U.- 201607-03).

## Disclosure

The authors report no conflicts of interest in this work

## References


1. Baartman LKJ, Quinlan KM. Assessment and feedback in higher education reimagined: using programmatic assessment to transform higher education. *Perspectives*. 2024;28(2):57–67.
2. Ginsburg S, Watling CJ, Schumacher DJ, Gingerich A. Numbers Encapsulate, Words elaborate: toward the best use of comments for assessment and feedback on entrustment ratings. *Acad Med*. 2021;96(7S):S81–S86. doi:10.1097/ACM.0000000000004089
3. Richardson D, Landreville JM, Trier J, et al. Coaching in competence by design: a new model of coaching in the moment and coaching over time to support large scale implementation. *Persp Med Educ*. 2024;13(1):33–43. doi:10.5334/pme.959
4. Lawrence K, van der Goes T, Crichton T, et al. Continuous reflective assessment for training (CRAFT): a national programmatic assessment model for family medicine; 2018. Avaiable from: https://www.cfpc.ca/CFPC/media/Resources/Faculty-Development/CRAFT_ENG_Final_Aug27.pdf. Accessed May 27, 2024.
5. Ginsburg S, Gingerich A, Kogan JR, Watling CJ, Eva K. Idiosyncrasy in assessment comments: do faculty have distinct writing styles when completing in-training evaluation reports? *Acad Med*. 2020;95:S81–S88.
6. Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. *Advan Health Sci Educ*. 2013;18(3):325–341. doi:10.1007/s10459-012-9372-1
7. Bogo M, Regehr C, Woodford M, Hughes J, Power R, Regehr G. Beyond competencies: field instructors' descriptions of student performance. *J Soc Work Educ*. 2006;42(3):579–593. doi:10.5175/JSWE.2006.200404145
8. Yilmaz Y, Jurado Nunez A, Ariaeinejad A, Lee M, Sherbino J, Chan TM-Y. Harnessing natural language processing to support decisions around workplace-based assessment: machine learning study of competency-based medical education. *JMIR Med Educ*. 2022;8(2):e30537. doi:10.2196/30537
9. Kelleher M, Kinnear B, Sall DR, et al. Warnings in early narrative assessment that might predict performance in residency: signal from an internal medicine residency program. *Perspect Med Educ*. 2021;10(6):334–340. doi:10.1007/S40037-021-00681-W
10. Maimone C, Dolan BM, Green MM, Sanguino SM, Garcia PM, O'Brien CL. Utilizing natural language processing of narrative feedback to develop a predictive model of pre-clerkship performance: lessons learned. *Perspec Med Educ*. 2023;12(1):141–148. doi:10.5334/pme.40
11. Van Ostaeyen S, De Langhe L, De Clercq O, Embo M, Schellens T, Valcke M. Automating the Identification of feedback quality criteria and the CanMEDS roles in written feedback comments using natural language processing. *Perspect Med Educ*. 2023;12(1):540–549. doi:10.5334/pme.1056
12. Moonen-van Loon JMW, Govaerts M, Donkers J, van Rosmalen P. Towards automatic interpretation of narrative feedback in competency-based portfolios. *IEEE Transact Learn Technol*. 2022;15(2):179–189. doi:10.1109/TLT.2022.3159334
13. Nair BR, Moonen-van Loon JMW, Parvathy M, Jolly BC, van der Vleuten CPM. Composite reliability of workplace-based assessment of international medical graduates. *Med J Aust*. 2017;207(10):453. doi:10.5694/mja17.00130
14. Nair BR, Moonen-van Loon JMW, Parvathy M, van der Vleuten CPM. Composite reliability of workplace based assessment of international medical graduates. *MedEdPublish*. 2021;10(1). doi:10.15694/mep.2021.000104.1
15. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res*. 2003;3(1):1.
16. Silge J, Robinson D. tidytext: text mining and analysis using tidy data principles in R. *J Open Source Software*. 2016;1(3). doi:10.21105/joss.00037
17. Gibbons C, Richards S, Valderas JM, Campbell J. Supervised machine learning algorithms can classify open-text feedback of doctor performance with human-level accuracy. *J Med Internet Res*. 2017;19(3):e65. doi:10.2196/jmir.6533

18. Dahal B, Kumar SAP, Li Z. Topic modeling and sentiment analysis of global climate change tweets. *Social Net Anal Mining*. 2019;9(24). doi:10.1007/s13278-019-0568-8
19. Müller W, Rebholz S, Libbrecht P. Automatic Inspection of E-Portfolios for Improving Formative and Summative Assessment. In: Wu TTGR, Huang YM, Xie H, Cao Y, editors. *Emerging Technologies for Education. SETE 2016. Lecture Notes in Computer Science*. Vol. 10108. Cham: Springer; 2017:480–489.
20. Boyd-Graber J, Hu Y, Mimno D. Applications of Topic Models. *Foundat Trend Inform Retriev*. 2017;11:143–296. doi:10.1561/1500000030
21. Grün B, Hornik K. topicmodels: an R package for fitting topic models. *J Statist Softw*. 2011;40(13):1–30. doi:10.18637/jss.v040.i13
22. Liu B. Sentiment Analysis and Subjectivity. In: Indurkhya N, Damerau FJ, editors. *Handbook of Natural Language Processing*. 2nd. 2010.
23. Pang B, Lee L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. *Proceed Assoc Comput Linguis*. 2004;42:271–278.
24. Pang B, Lee L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. *Proceed Assoc Comput Linguis*. 2005;43:115–124.
25. Pang B, Lee L. Opinion mining and sentiment analysis. *Foundat Tren Inform Ret*. 2008;2(1–2):1–135. doi:10.1561/1500000011
26. Cohen GS, Blumberg P, Ryan NC, Sullivan PL. Do final grades reflect written qualitative evaluations of student performance? *Teach Learn Med*. 1993;5(1):10–15. doi:10.1080/10401339309539580
27. Bartels J, Mooney CJ, Thompson Stone R. Numerical versus narrative: a comparison between methods to measure medical student performance during clinical clerkships. *Med Teach*. 2017;39(11):1154–1158. doi:10.1080/0142159X.2017.1368467
28. Eva KW, Hodges BD. Scylla or Charybdis? Can we navigate between objectification and judgement in assessment? *Med Educ*. 2012;46 (9):825–920. doi:10.1111/j.1365-2923.2012.04310.x
29. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. *Med Teach*. 2024;46 (4):446–470. doi:10.1080/0142159X.2024.2314198