Advances in Health Sciences Education (2022) 27:23–48
https://doi.org/10.1007/s10459-021-10067-6

Using theory-informed data science methods to trace the quality of dental student reflections over time

Yeonji Jung[1] · Alyssa Friend Wise[1] · Kenneth L. Allen[2]

Received: 10 December 2020 / Accepted: 22 August 2021 / Published online: 2 September 2021

Abstract

This study describes a theory-informed application of data science methods to analyze the quality of reflections made in a health professions education program over time. One thousand five hundred reflections written by a cohort of 369 dental students over 4 years of academic study were evaluated for an overall measure of reflection depth (No, Shallow, Deep) and the presence of six theoretically-indicated elements of reflection quality (Description, Analysis, Feeling, Perspective, Evaluation, Outcome). Machine learning models were then built to automatically detect these qualities based on linguistic features in the reflections. Results showed a dramatic increase from No to Shallow reflections from the start to end of year one (20% → 66%), but only a limited gradual rise in Deep reflections across all four years (2% → 26%). The presence of all six reflection elements increased over time, but inclusion of Feelings and Analysis remained relatively low even at the end of year four (found in 44% and 60% of reflections respectively). Models were able to reliably detect the presence of Description ($\kappa_{TEST}=0.70$) and Evaluation ($\kappa_{TEST}=0.65$) in reflections; models to detect the presence of Analysis ($\kappa_{TEST}=0.50$), Feelings ($\kappa_{TEST}=0.54$), and Perspectives ($\kappa_{TEST}=0.53$) showed moderate performance; the model to detect Outcomes suffered from overfitting ($\kappa_{TRAIN}=0.90$, $\kappa_{TEST}=0.53$). A classifier for overall depth built on the reflection elements showed moderate performance across all time periods ($\kappa_{TEST}>0.60$) but relied almost exclusively on the presence of Description. Implications for the conceptualization of reflection quality and providing personalized learning support to help students develop reflective skills are discussed.

Keywords Reflection · Health professions education · Educational data sciences · Classification

✉ Yeonji Jung
yeonji.jung@nyu.edu

Alyssa Friend Wise
alyssa.wise@nyu.edu

Kenneth L. Allen
kenneth.allen@nyu.edu

[1] Learning Analytics Research Network (NYU-LEARN), New York University, 370 Jay Street, 5th Floor, Brooklyn, NY 11201, USA

[2] Department of Cariology and Comprehensive Care, College of Dentistry, New York University, 137 E. 25th Street, 6th Floor, New York, NY 10010, USA

Background

Reflection is a critical skill in health professions education that can support students in becoming thoughtful practitioners and life-long learners (Schon, 1983). Reflective practice offers a powerful way for students and practitioners to competently address an ever-expanding knowledge base and continuously improve their skills to provide a better quality of care (Mann et al., 2009). Despite widespread inclusion of reflection in training for health professions, students are rarely meaningfully assessed or given specific feedback to help them develop reflective skills (Koole et al., 2011). This is due to both the time-consuming nature of manual assessment and the common use of simplistic quality criteria that do not offer detailed information for improvement. It is thus an area in which there is potential value in applying educational data science approaches to generate feedback on reflections at scale (Ullmann, 2019). In doing so, it is important to both draw on existing conceptualizations of reflection quality and consider how the examination of these constructs in large data sets can be used to validate and refine them, addressing recent critiques of a lack of theory-informed work applying data science methods to health professions education (Tolsgaard et al., 2020).

This study approaches these issues by (a) examining the presence of, and relationships between, six elements of reflection quality (Description, Analysis, Feeling, Perspective, Evaluation, Outcome) and an overall measure of reflection depth (No, Shallow, Deep) in the professional development statements of 369 dental students over their four years of academic study; and (b) building and evaluating the performance of machine learning models to automatically detect these qualities based on linguistic features in the reflections. The findings offer insights into the nature of dental students' reflections and represent a first step towards the creation of systems for personalized reflection support.

The importance of reflection in health professions education

Reflection is widely defined as a careful consideration of one's experiences used to make strategic changes for the future (Boud et al., 1985; Dewey, 1933). In professional education, it is often connected to the notion of a "reflective practitioner" (Schon, 1983), one who maintains awareness of how they conduct their professional practice and adjusts it in the moment as needed. An important stepping stone to such reflection-*in*-action is reflection-*on*-action: being able to think back on one's prior professional experiences, identify gaps between actual and desired practice, and develop strategies to address them (Mann et al., 2009). For these reasons, reflective writing is widely used in health professions education through required essays, journaling, or portfolios (e.g. Heeneman & Driessen, 2017). These can be useful both for students to identify current learning needs and develop reflective skills for later use as a professional (Mann et al., 2009). However, the ability to reflect effectively on one's learning does not develop automatically (Bush & Bissell, 2008). Students need feedback that evaluates actual reflections against some criteria for quality, a process referred to as reflection assessment (Heeneman & Driessen, 2017).

Prior work evaluating the quality of reflection

The majority of prior efforts to study the quality of health professions reflections have employed unidimensional measures of reflection quality, indicating an underlying belief that quality varies along a single scale. The most common approach draws on Mezirow

(1991)'s three-level scheme of No, Shallow or Deep (Critical) reflection. For example, Tsingos et al. (2015) adapted this scheme to appraise pharmacy students' reflections. In this work, they suggested that the levels of reflection differ in the specificity of experiences described and their connection to lessons learned, ranging from no description of experiences (No Reflection), to presentation of experience (Shallow Reflection), to presentation with implications (Deep Reflection).

A variety of research has been conducted using such unidimensional criteria to manually assess depth in student reflections. On the whole, research findings document that reflection generally occurs less frequently than desired, and it is more often shallow than deep (Hanson & Alexander, 2010; Wong et al., 1995). Other work has shown that not only does reflection occur less deeply than desired, but that students do not necessarily improve their reflections on their own over time (Moon, 2013). This underlines the need to provide students with feedback; however simplistic feedback that a reflection is of "low depth" does not offer direction on how to improve and thus can be perceived as less-than-meaningful to students (Koole et al., 2011), at times triggering negative reactions and actually causing them to engage less (Bush & Bissell, 2008).

Compounding the issue is the fact that manual assessment of reflection is time-consuming (Koole et al., 2011). This has led to an interest in the use of educational data science approaches for formative reflection assessment (Kovanovic et al., 2018). Initial automated approaches to reflection assessment use computational approaches to build models that classify reflections into one of several levels of quality along a single scale. For example, Liu et al. (2019) developed a random forest classifier that showed good performance (F-score = 0.80) for the binary classification of whether or not reflection occurred in texts pharmacy students wrote about their work placements. Predications were made based on the presence of various linguistic features extracted from the Linguistic Inquiry and Word Count (LIWC) and the Academic Writing Analytics platforms. Outside health professions education, Kovanovic et al. (2018) developed a random forest classifier for undergraduate students' reflections across multiple disciplines. Their model used lexical dictionaries (LIWC and Coh-metrix) and ngrams (strings of words) to classify reflections as achieving one of three levels of quality according to the specificity of goals (Observation < Motive < Goal) with moderate performance (accuracy = 0.75, $\kappa = 0.51$). In their model the Observation and Motive levels of reflection were highly indicated by the use of past-oriented words and the highest Goal level was associated with causal and perceptual words. These prior studies suggest linguistic features can provide a good basis for automated reflection assessment; however such unidimensional models indicating only "more" or "less" depth still offer a relatively limited basis for providing constructive feedback to students.

An alternative approach to understanding reflection quality uses a multidimensional framing, suggesting the presence of multiple distinct elements that contribute to quality. The earliest version of these positioned the elements as hierarchical, adding specificity but still only offering a single linear path available for improvement (e.g. Attending to Feelings → Association → Integration → Validation → Appropriation → Outcome, Boud et al., 1985). However, other formulations took the elements to be relatively independent, with overall depth resulting from their collective presence or absence. For example, Gibbs (1988) reformulated Boud's framework as an expanded set of elements that could occur in different combinations: Description, Feelings, Evaluation, Analysis, Conclusion, and Action Plan. There has been recent interest in such multidimensional conceptualizations as a way to more robustly appraise reflection quality and offer actionable guidance for its improvement. In health professions education, Cui et al. (2019) applied this approach to develop a framework of six reflection elements for dental education based on a synthesis of different schemes found in the literature: Description, Analysis, Feelings, Perspective, Evaluation, and Outcome. These elements were indicated for computational assessment as a succinct set of conceptually distinct entities aligned with possible linguistic features. While empirical work documenting differences in the prevalence of these features across reflections was promising, Cui et al. (2019) stopped short of training a classifier based on the labeled data.

Outside health professions education, there have been some attempts at automated multi-dimensional reflection assessment. Ullmann (2019) developed multiple binary classifiers for undergraduate students' reflections across various subjects. The models used unigrams to classify the presence of eight elements with models showing the most substantial performance for the Experience element and the lowest performance for the Perspective element. Ullmann's work offers proof-of-concept that automated multidimensional assessment can be usefully applied to student reflections; however, it has several limitations. First, the different classifiers were built separately, ignoring potential information about relationships between elements. Second, the unit of analysis was set at the sentence level, limiting detection of elements expressed across multiple phrases (Moon, 2013). Finally, using specific unigrams (words) from the text as predictive features limits the model's applicability to other health professions education contexts. In another effort, Gibson and colleagues (Gibson et al., 2016, 2017) worked with undergraduate students from diverse disciplines to develop a multidimensional automated reflection assessment tool. In this work, elements of reflection quality were conceptualized as aspects of metacognition and predicted using parts-of-speech (POS) patterns. While their work did not examine the details of how the different elements occurred in the reflections, the cumulative presence of the metacognitive elements was useful to predict overall reflection quality as weak or strong (accuracy > 0.75). This work added to the evidence base by suggesting a potential relationship between the elements-based approach and overall reflection quality.

The current study

Prior work indicates a need to provide students with specific (multi-dimensional) feedback in order to help them progress from simplistic forms of reflection to more advanced ones (Chirema, 2007; Hanson & Alexander, 2010; Koole et al., 2011). Further, the time-consuming nature of manual assessment presents an opportunity to use data science methods to build machine learning models that can automate this process. Initial efforts document the viability of detecting the overall level of reflection (Kovanovic et al., 2018; Liu et al., 2019) and presence of specific reflective elements (Cui et al., 2019; Ullman, 2019) based on language use, as well as suggesting a relationship between the two (Gibson et al., 2016, 2017). However, no work has empirically probed the nature of this relationship, nor built a generalizable model to provide feedback based on it. The current study fills this gap by (a) examining the presence of and relationships between six elements of reflection quality and an overall measure of reflection depth in the professional development statements of 369 dental students over their four years of academic study; and (b) building and evaluating the performance of machine learning models to automatically detect these qualities based on linguistic features in the reflections.

Addressing recent calls to use theory to inform the application of data science methods to health professions education (Tolsgaard et al., 2020), this study employs two existing conceptualizations of reflection quality: (1) Overall reflection depth based on Tsingos et al.'s (2015) widely used framework adapting Mezirow (1991)'s three-level scheme (No, Shallow, and Deep Reflection) for the context of health professions education; (2) Reflection elements based on Cui's et al. (2019) framework synthesizing the most commonly discussed aspects of reflection in health professions education into a set of six elements (Description, Analysis, Feeling, Perspective, Evaluation, and Outcome) aligned with linguistic data features.

The findings from this work both offer insights into the nature of dental students' reflections and represent a first step towards building systems for personalized reflection support. The specific research questions are as follows:

RQ1 What reflection elements are present in dental students' reflections and how do they change over time?
RQ2 Can the presence of reflection elements be predicted using linguistic features and how does model performance vary across different time periods?
RQ3 What are overall levels of reflection depth and how do they change over time?
RQ4 Can the overall level of reflection depth be predicted using linguistic features/the presence of reflection elements and how does model performance vary across different time periods?

Methods

Context of learning and reflection

Students in a US four-year pre-doctorate dental program are required to take a self- and peer-assessment skills course every academic year. In years 1 and 2 assessment skills exists as a stand-alone graded course (with the self- and peer-assessment components given equal weighting); in years 3 and 4 it is integrated with a comprehensive patient care course and worth 20% of the total grade. In all years the self-assessment component of the course guides students to respond to various reflective prompts through a custom online e-Portfolio system. This system was designed as a mentored environment for students as future dentistry practitioners to develop self-assessment skills and knowledge through reflective writing. While some reflection prompts asked about specific content (e.g. courses or competencies), this study focused on overall reflection statements that students were required to complete at the start and end of every academic year, asking for "thoughtful personal reflections on your goals, the current state of your knowledge and competence and the successes and challenges you have encountered or anticipate in your education." The statement categories were personal (public/private), becoming a professional (public/private), ethics (public), and professional progress (public). Public statements were available to be viewed by peers in the same practice group, while private statements were accessible only to the student and faculty mentor assigned to the group. The written guidelines intentionally left room for interpretation and discussion with the faculty mentor who guided students' reflective writing. Faculty mentors additionally reviewed the reflections over time, reaching out to students through the system as needed. Faculty mentors in years 3 and 4 are Group Practice Directors who know the students well, are aware of their clinical expertise/patient interactions and can substantively comment on the student entries. Reflections were graded for their timeliness and quality with respect to the expectations described above; no additional rubric was provided.

Reflection corpus & participants

The initial data corpus consisted of all 7510 reflections submitted by the 378 students in a single graduating class, categorized into one of eight time periods corresponding to the first and second half of each of the four years (e.g. D1 Start indicates first semester of the first year; D4 End indicates the last semester of the fourth year). Removal of 73 reflections with no values, 2572 with duplicated contents, and 292 outliers (length $\pm$ 3 SD) yielded a final corpus of 4573 reflections. Stratified random sampling across the eight time periods and six reflection types yielded a final representative set of 1500 reflections from 369 students. Reflections, on average, contained 110 words and 5 sentences.

Manually coding reflections for elements and depth (content analysis)

Coding schemes to assess the presence of each reflection element (based on the conceptual framework of Cui et al., 2019) and overall depth (No/Shallow/Deep reflection, Tsingos et al., 2015) were constructed iteratively using sample data not included in the study. One substantive change was made to the scheme of Cui et al. (2019): Perspective was clarified as referring to taking into account the views of others, making changes in one's own perspective to be considered as an Outcome (as part of lessons learned). Coding schemes included detailed descriptions and multiple examples (see Tables 1 and 2 for abridged versions).

The entire text of a reflection was taken as the unit of analysis since elements and depth can occur across sentences (Moon, 2013) and reflections were generally relatively short. Coder training was conducted by two researchers on sample data not included in the study until reliability was stable at an acceptable level ($\alpha > 0.70$ for all seven judgements; Artstein & Poesio, 2008) using Krippendorff's unweighted α calculated separately for each of the six binary elements and the weighted version of α for the three-level depth coding. The minimum proportion of reflections to double-code was calculated as 30% based on the results of coder training, using Cantor's (1996) method to infer the desired reliability level for the entire coded sample ($\alpha > 0.70$, p-value $= 0.05$, Power $= 0.80$). A final proportion of 33% (500 reflections) was double coded at even intervals of 125 reflections each across seven rounds of coding. Inter-rater reliability was good for each of the six elements ($\alpha_{overall} > 0.75$, $\alpha_{round} > 0.70$) and depth ($\alpha_{overall} = 0.75$, $\alpha_{round} > 0.67$). Disagreements between the coders were reconciled through discussion until consensus was reached. Chi-square tests were used to identify differences in the presence of elements and level of depth across time periods and strength of association among elements and with depth was assessed using Cramer's V.

Computationally extracting linguistic features from the reflections

76 of the 93 linguistic features in LIWC 2015 (Pennebaker et al., 2015) were extracted from the reflections; the remaining seventeen features were excluded due to redundancy. Use of a pre-defined dictionary supports model generalizability to other contexts (as compared to the extraction of common words) and LIWC features have shown good performance for classifying reflections in similar higher education contexts (Kovanovic et al., 2018; Liu et al., 2019). In addition, LIWC's theoretical construction of linguistic features

Table 1 Abridged coding scheme for six reflection elements

Elements	Definition	Example
Description	Students indicate what specific experience happened to them	I had to leave my family and friends and come to a country where I had no one but my husband
Analysis	Students explain the causes or consequences of their experiences	*Causes* This difficulty I suffered from can be for multiple reasons; for one, there are often shortcuts in didactic courses that can help one achieve high grades easily
		Consequences Everything I learned at school will help/allow/lead me to develop the skills to become a confident dentist
Feeling	Students describe the emotional reaction of how they felt before, during, or after the experience	Looking back now, I am surprised and proud of myself for the amount of work I have been able to go through
Perspective	Students describe a consideration of others' situations, feelings, needs, or intentions	It's important to listen to a patient's complaints and be cognizant of their comfort within a clinical setting
Evaluation	Students make a judgment of what was good or bad about the experience	Competing with many highly qualified dentists who applied from all over the world was extremely difficult
Outcome	Students describe lessons learned from the experience or future intentions made based on experience	*Lessons learned* I made up my final decision that grades are not the most important aspect in here
		Future intention/resolution In the next 2 years in the clinic, there will be opportunities for me to gain hands-on experiences

Table 2 Abridged coding scheme for reflection depth

Levels	Definition	Example
Deep Reflection	Students demonstrate the experience specified with clear links to its lesson	The lessons I have learned have been greater than that in academia; The first lesson I have is that failure is okay. Secondly, I am figuring out I need to let things go sometimes. If something goes wrong, it's not always because of my incompetence
Shallow Reflection	Students attempt to specify the experiences without its implications to them	I feel like as I'm learning more and more from school I have more responsibilities to take good care of my patients
No Reflection	Students describe hope or objective facts without any experience specified	As a healthcare professional, I should treat my patients like I would want to be treated

supports model interpretation. While LIWC does not use part-of-speech tagging or negation, prior work has found only a modest resulting reduction in model performance (Crossley et al., 2017).

Building classifiers to predict the presence of reflection elements based on linguistic features

Two different approaches were tested to predict the six reflection elements based on the presence of LIWC linguistic features: one in which the presence of each of the reflective elements was predicted independently and one in which the presence of all elements was predicted simultaneously, taking into account potential relations between presence of the different elements (Herrera et al., 2016).

To predict the presence of each of the reflective elements independently, six single-label classifiers were trained using the caret package for R. Multiple classification algorithms were tested by building models on a training set which consisted of 80% of the data and then evaluating using ten-fold cross-validation (a technique that randomly partitions the input data into 10 subgroups, trains the model using 9 subgroups, validates the model using the held-out subgroup, and then repeats this process 10 times so that each subgroup is held-out once, averaging overall performance across the iterations). An additional evaluation was conducted by assessing each model's performance on the hold-out test set of 20% of the data (James et al., 2013).

Results followed previous findings showing random forest models to outperform other methods in similar reflection classification tasks (Liu et al., 2019). Random forest is an ensemble classification technique that provides low-bias, low-variance performance and allows for inspection of feature importance by constructing multiple decision trees using random subsets of the features on bootstrapped samples and making the classification decision based on a majority voting mechanism (Breiman, 2001). Random forest model optimization was conducted on the training set through specification of two hyperparameters: ntree (the number of trees in the ensemble) and mtry (the number of random features tested at each branch of the tree). Performance of all six classifiers stabilized at ntree = 500 (comparing error rates at 100 tree intervals), and best performance was achieved using mtry = 32 (based on the default grid search strategy). Optimized models were evaluated using ten-fold cross-validation and on the 20% hold-out test set (both overall and divided by year to assess temporal variability in performance). Feature importance was assessed using the Mean Decrease Gini (MDG) index, which measures the average decrease of a given feature in Gini impurity across all tree nodes (James et al., 2013).

While the approach described above created six independent classifiers that each assigned a single binary label to a reflection (e.g. Analysis or No Analysis), a multi-label classifier can assign multiple labels to each reflection simultaneously, taking into account any potential relationships between them (e.g. the presence of Analysis and Evaluation may not be independent; Herrera et al., 2016). To apply this approach, first, the multi-label problem is transformed into a set of single-label problems. This can be done in a simple manner via the binary relevance method (which creates a separate yes/no classification task for each label) or in more complex ways such as the classifier chain method (in which the attribute space for each binary model is extended using the 0/1 label relevances of all previous classifiers; Read et al., 2011). Similar to above, multiple transformation algorithms were tested on the training data. In this case, models were built using Meka (a multi-target extension application to the Weka machine learning software), and performance was evaluated using the macro-averaging method (which measures the average value of evaluation metrics for each label classification over the total number of labels). Results showed that classifier chains (CC) outperformed all other methods, showing the highest likelihood that all labels are classified correctly (exact match score) and lowest number of labels likely to be incorrectly predicted (Hamming loss score). The final CC classifier was assessed using ten-fold cross-validation and on the 20% hold-out test set. Feature importance was measured using average Chi-squared estimates of each feature across all the elements.

Building classifiers to predict reflective depth based on linguistic features or reflective elements

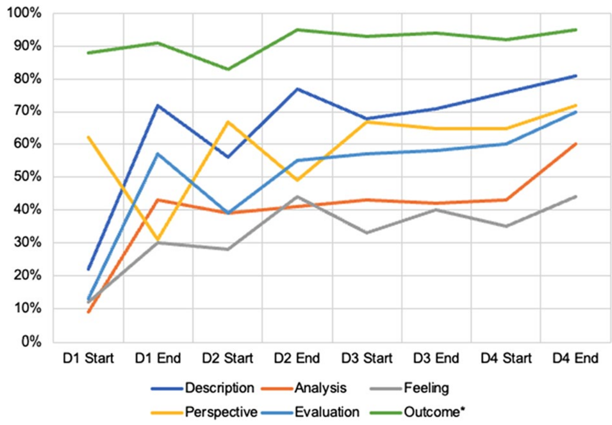
Reflection depth was modeled using random forest classifiers built on the 80% training data using two different feature sets. First, a model for reflection depth was trained directly on LIWC linguistic features as described above. Second, a model was trained based on six features set as the presence/absence of each reflection element. This was done both to probe the relationship between depth and elements, as well as to examine the impact on the overall accuracy of depth prediction. The manual codes were taken as the most reliable measure of ground truth available and used as a basis for model training. Model testing on the 20% held-out test set was conducted using both the manually-coded values and the predicted values from the first set of models to assess relative performance and viability of automated prediction of reflection depth. In both cases, random forest classifiers were trained, optimizing ntree and mtry as described previously. Classification performance stabilized at 500 trees for both classifiers. For the model using linguistic features, best performance was achieved using mtry = 32; for the model using reflection elements as features, all of the 6 features were available for use at each of the branches. Optimized models for both cases were evaluated using ten-fold cross-validation and on the 20% hold-out test set (both overall and divided by year to assess temporal variability in performance). The model using reflection elements as features was tested using both manually coded and predicted values for elements. Feature importance for both models was assessed using the MDG index.

Results

RQ1: what reflection elements are present in dental students' reflections and how do they change over time?

Description (70%), Perspective (65%) and Evaluation (57%) were found in the majority of reflections, while Analysis (45%) and Feeling (37%) occurred less frequently. Outcome was included in virtually all reflections (93%). Associations between the elements varied. While Perspective and Outcome had a low association with the other elements ($0.00 < V < 0.26$), Description, Analysis and Evaluation had strong associations with each other ($0.57 < V < 0.76$). Feeling showed moderate to strong association with Description, Analysis and Evaluation ($0.34 < V < 0.50$). These levels of association indicate that it is important to consider the relationship between elements in performing the classification tasks (see RQ2). Changes in the presence of elements over time were found using Chi-squared tests. First, D1 Start (first semester of the first year) showed dramatically lower levels than any of other time periods for most elements (except Perspective and Outcome, see Fig. 1). Second, D4 End (last semester of the fourth year) showed a higher presence

Fig. 1 Presence of reflection elements over time. *Before class rebalancing using SMOTE



of all elements than prior time periods, but to varying degrees. Looking more specifically, Description, Analysis, and Evaluation showed a notable rise in both D1 and D4 from the start to end of the year, but relatively consistent levels in between. For Feeling, levels fluctuated within a consistent range after the initial rise following D1 Start. Perspective was relatively consistent throughout, except for a notable dip in D1 End and small rise in D4 End. Outcome was present in high levels across all time periods. The varying levels of element presence indicate the importance of developing classifiers that can offer good performance on data from different years.

RQ2: can the presence of reflection elements be predicted using linguistic features and how does model performance vary across different time periods?

Prior to building the classifiers, the data was examined for modeling suitability. A very high percentage of reflections containing Outcome (93%) created the problem of class imbalance (not enough cases of one class (No Outcome) from which to build a model). This issue was addressed using SMOTE (Synthetic Minority Oversampling Technique; Chawla et al., 2002) which creates synthetic instances of the underrepresented class (No Outcome) to balance the data. The final proportion of data used for all modeling had 34% of cases with No Outcome (enhanced from the naturally occurring 7%).

The optimized single label classification models for reflection elements showed moderate ($\kappa > 0.40$) to substantial ($\kappa > 0.60$) performance on both training and test sets (Landis & Koch, 1977) (see Appendix Table 3 for full model performance statistics). Description and Evaluation classifiers showed the best performance ($\kappa \geqq 0.60$) with recall and precision both above 80%. The Perspective classifier had high accuracy ($\geqq 0.75$), but reduced kappa ($\geqq 0.44$), while Analysis and Feeling classifiers had moderate kappa ($\geqq 0.50$), with higher precision ($\geqq 0.74$) than recall ($\geqq 0.59$), indicating that the presence of these elements was missed by the classifiers in some cases. This was particularly notable for Feeling which had a recall of only 60%. The Outcome classifier showed good performance in cross-validation ($\kappa = 0.90$) but a dramatic decrease in performance on the test set ($\kappa = 0.53$) suggesting problems with overfitting. Overall, while there is room for improvement, results indicate that linguistic modeling can be useful to assess the presence of individual reflection elements using single-label classifiers.

For the single-label classifiers, Description, Analysis, and Evaluation classifiers all had skewed distributions of features indexed by the MDG scores, indicating that model performance depended heavily on a small number of features (see Appendix Table 9 for details). These three reflection elements were most strongly predicted by words related to past events (*LIWC feature: focuspast*). Description and Evaluation were next strongly predicted by words expressing personal voice (*LIWC feature: authentic*) while Analysis depended more on text length (*LIWC feature: word count*). After this, the three element classifiers showed similar patterns of feature importance including the presence of words indicating orientation to time (*LIWC feature: time*), comparisons to an ideal status (*LIWC feature: discrepancy*, e.g. should, would), and the use of quantifiers (*LIWC feature: quant*, e.g. few, much). Description and Evaluation were also predicted by the use of the first-person pronoun (*LIWC feature: i*) while Analysis was predicted by words showing causal inferences (*LIWC feature: cause*) or thoughtful ideas (*LIWC feature: insight*). The Outcome classifier also showed some skew in MDG, with use of the first-person pronoun (*LIWC feature: i*) as the most important feature, followed by features related to social relations (*LIWC features: clout, affiliation, we, you*), use of articles (*LIWC feature: article*), expression of personal voice (*LIWC feature: authentic)* and words referring to events to come (*LIWC feature: focusfuture*). In contrast, Feeling and Perspective classifiers had relatively flat distributions, indicating that multiple features were important for the classification tasks. While showing some similar predictive features to the first three classifiers (*LIWC features: focuspast, word count, authentic*), Feeling reflections included language related to compensation (*LIWC feature: reward*) and emotional words including perceptions (*LIWC feature: feel*), negative emotions (*LIWC feature: anxiety*) and positive emotions (*LIWC feature: posemo*). On the contrary, for Perspective reflections, the use of third-person pronouns (*LIWC feature: they*) was the most important feature, followed by expressions of personal voice (*LIWC feature: authentic*), positive emotions (*LIWC feature: posemo*), indications of temporality (*LIWC feature: time*), and focus on tasks (*LIWC feature: work*).

The results for the multi-label classification model showed an overall accuracy of 0.80, recall of 0.79, precision of 0.78, and exact match of 0.31. Among the individual labels assigned based on the classification, Outcome showed the highest recall and precision (both > 0.90), while Description, Perspective, and Evaluation also showed relatively high recall (> 0.85) and precision (> 0.75). Similar to the single-label models, Analysis and Feeling showed lower recall than other elements (0.66 and 0.53 respectively), indicating that many instances of Analysis and Feelings in reflections were undetected. In alignment with the single-label classifiers, the multi-label classifier also showed a skewed distribution of feature importance, indicating heavy reliance on a small number of features (mean feature importance = 0.038, SD = 0.065, max = 0.271). In the multi-label model, personal voice (*LIWC features: authentic, I*) and time orientation (*LIWC features: focuspast, time*) were most predictive of the elements, followed by the text length (*LIWC feature: word count*), display of social relations (*LIWC feature: clout*), and indications of feelings (*LIWC feature: feel*). Words related to cognitive processes such as making comparisons (*LIWC feature: discrepancy*), numerical comparisons (*LIWC feature: quant*), and analytical thinking (*LIWC feature: analytic*) were also useful in predicting the elements.

Regarding the variability in prediction performance across different time periods, single-label models for Description, Feeling, Evaluation and Outcome showed relatively consistent model performance across time periods with two exceptions (see Fig. 2). For Description, the model showed a reduced kappa in D2, and for Evaluation the model had reduced precision (though high recall) in D1. Models for Perspective and Analysis showed more variation in performance across time periods generally, with the Analysis model

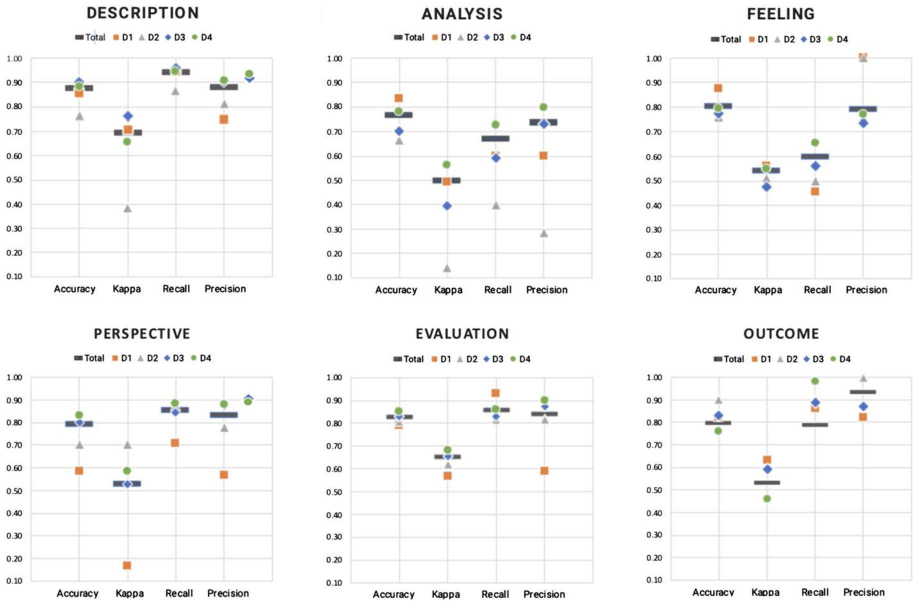


Fig. 2 Model performance of the reflection elements classifiers across time periods

performing particularly poorly in D2 and the Perspective model performing poorly in D1. Because the multi-label model did not show overall improved performance over the single-label models, its temporal variability was not tested.

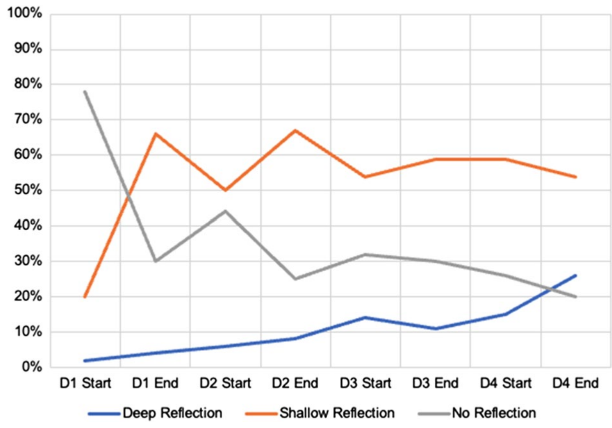
RQ3: what are overall levels of reflection depth and how do they change over time?

Overall depth of reflection was found to most frequently be Shallow (53%) or No Reflection (31%), while Deep Reflection was far less common, occurring in only 16% of the time. Changes in the relative proportion of each level of reflective depth were identified using Chi-squared tests. Depth of reflection showed a dramatic shift from the start to end of D1 with the proportion of No Reflection dropping from 78 to 30% and Shallow Reflection rising from 20 to 66% (see Fig. 3). There was also a slow rise in Deep Reflection from D1 Start (2%) through D2 End (8%), with greater increases by D3 Start (14%) and D4 End (26%).

RQ4: can the overall level of reflection depth be predicted using linguistic features/the presence of reflection elements and how does model performance vary across different time periods?

Performance of the optimized model for reflection depth based on linguistic features showed moderate kappa on the training data ($\kappa = 0.55$) and the holdout test data ($\kappa = 0.63$), with recall and precision ranging from 0.70 to 0.80 (see Appendix Table 4 for full model performance statistics). The Shallow class had the highest recall (0.85 compared to 0.72/0.71) with cases of Shallow reflection misclassified as Deep or No Reflection less

Fig. 3 Levels of reflection depth over time



frequently than the reverse. There was no confusion between cases of No and Deep Reflection (see Appendix Table 5 for the complete confusion matrix).

A small number of linguistic features played an important role in model performance (mean MDG = 0.07, SD = 16, max = 102.12; see Appendix Table 10 for details). Similar to the models for Description, Analysis and Evaluation (see RQ2), the most important features were related to use of past-oriented words (*LIWC feature: focuspast*), the text length (*LIWC feature: word count*), and expressions of personal voices (*LIWC feature: authentic*), with these elements being most present in Deep Reflections and more present in Shallow Reflections than those which contained No Reflection. Writing showing No Reflection had a relatively high use of future-oriented (*LIWC feature: focusfuture*) and comparisons words (*LIWC feature: discrepancy*), whereas both Shallow and Deep Reflections contained more use of the first-person pronoun (*LIWC feature: i*), present-oriented words (*LIWC feature: focuspresent*), temporality (*LIWC feature: time*), and quantifiers (*LIWC feature: quant*). Shallow Reflections showed a greater emphasis on current events (*LIWC feature: focuspresent*) than the other two classes.

Prior to building the classifier of reflection depth based on reflection elements, associations between depth and the six elements were tested; Depth was highly associated with Description (V = 0.98), followed by Evaluation (V = 0.76) and Analysis (V = 0.64), but had only moderate association with Feeling (V = 0.50) and low association with Perspective (V = 0.16) and Outcome (V = 0.26). Performance of the optimized model built on the manually coded elements was better than that of the model built directly on the linguistic features (accuracy > 0.80, $\kappa > 0.70$, see Appendix Table 6 for full model performance statistics). Recall was good for the No and Shallow Reflections; however, Deep Reflections were commonly misclassified as Shallow. There was no confusion between No and Deep Reflections and little confusion between No and Shallow Reflections (see Appendix Table 7 for the complete confusion matrix). Testing the same model, but inputting values for element presence produced computationally by the six single-label classifiers showed reduced reliability compared to use of the manually coded values (accuracy = 0.73, $\kappa = 0.53$) but comparable performance to the model built directly on the linguistic features. The pattern of confusion was similar to that found when manually coded values were used (see Appendix Table 8).

Looking at the weight of the six elements in the model, Description was the by far most important feature, serving as virtually a binary divider between Shallow/Deep and

No Reflection. Analysis, Evaluation and Perspective were also absent from writing showing No Reflections, and present in greater amounts in Deep than Shallow reflections (see Appendix Table 11 for details). Feeling was more present in Deep and No Reflection than Shallow Reflections, while Outcome was not a good predictor, being relatively equally present in all depths of reflection.

Regarding the variability in prediction performance across different time periods, the depth classifier built directly on linguistic features showed relatively consistent performance across time periods except for D2, in which kappa was notably low ($\kappa=0.22$) (see Fig. 4a). This is partly attributed to the small size and imbalance of the D2 test set (7 No reflections, 14 Shallow reflections, and 0 Deep reflection). The depth classifier built on reflection elements showed relatively consistent performance across time periods when the manually coded element values were input (see Fig. 4b), but less temporal stability when predicted values for the elements were used, particularly in D2 ($\kappa=0.16$) (see Fig. 4c), repeating the pattern seen for predictions of Description (see Fig. 2).

Discussion

Summary of key results

Working with a corpus of reflections made by a class of 369 dental students over the four years of their program, this study examined the presence of different measures of reflection quality, their associations and changes over time, as well as the performance of classifiers built to detect their presence. Results showed a dramatic increase from No to Shallow Reflection from the start to end of year one, but only a limited gradual rise in Deep Reflection across all four years. The presence of all six reflection elements increased over time, but inclusion of Feelings and Analysis remained relatively low even at the end of year four. Classifiers were able to reliably detect Description and Evaluation in most time periods; classifiers for Analysis, Feeling, and Perspective showed moderate performance with room for improvement; the classifier for Outcome suffered from overfitting. Associations between the elements led to similarities in predictive feature importance across models, especially heavy reliance on the use of past-oriented words in classifiers for Description, Analysis and Evaluation. Multi-label classification of elements did not show substantial performance improvement over the single-label models. The Depth classifier built on reflection elements gave moderate performance across all time periods using manually

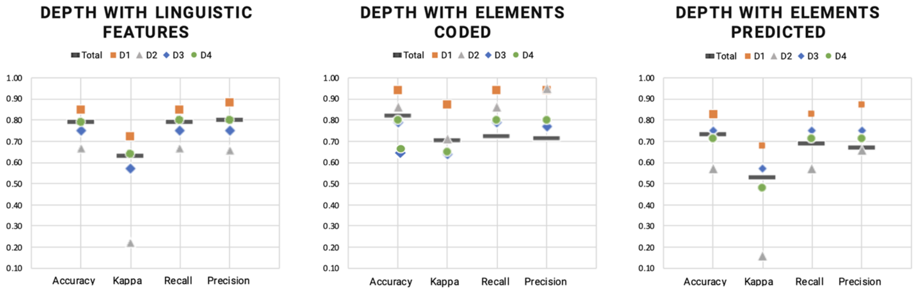


Fig. 4 Model performance of the reflection depth classifier across time periods

coded element values; errors were due primarily to confusion between Deep and Shallow reflections. The model's prediction was based almost exclusively on the presence of Description; thus applying it using predicted values for the elements produced a similar performance pattern to the Description model. Finally, the Depth classifier built directly on linguistic features showed moderate performance except for in D2, with errors due to misclassification of Shallow reflections as Deep or No-Reflection. Use of past-oriented words was again the strongest linguistic predictor.

Student development of reflective skills over time and needs for support

Following the start of their first year, the majority of students did reflect; however by the end of four years only a quarter of all statements showed deep, rather than shallow reflection. This aligns with prior findings that health professions students are not always fully aware of what they are learning (Chirema, 2007; Hanson & Alexander, 2010). While they may become somewhat more aware of their academic life as they become accustomed to the course difficulty and overwhelming workload (Alzahem et al., 2011), these results underline the fact that most students will not naturally develop deep reflection skills without explicit support.

Looking at the development of reflection elements, most showed a dramatic increase after the start of the first year and then remained relatively stable until a small final rise at the end of the fourth. This suggests that the most appropriate time to assess reflection and offer support is not during the initial year, but subsequent to this naturally-occurring rise. While the exact timeline may differ across contexts, the general pattern of an unprompted improvement shortly after reflective activities begin is well-documented (Chirema, 2007; Ip et al., 2012). Applying reflection models similar to the ones developed here can help to identify when this improvement period occurs in other contexts, providing a concrete example of how educational data sciences can support more effective pedagogical decision-making.

Turning to specific elements, the 369 dental students studied included Description in the vast majority of their reflections after the start of their first year and Outcome was very highly present across all periods of time. In contrast, Analysis occurred in less than half of reflections and Feelings were included even less frequently across all time periods. This suggests students may not see these elements as relevant for reflection unless they are explicitly prompted to include them (Ullmann, 2019); in the case of Feelings they may also be less used to and comfortable writing about them. Despite the relative lack of Analysis, Evaluation occurred in over half the reflections; this is potentially problematic since Analysis is generally useful to inform Evaluation. Finally, even though Perspective was included in two-thirds of reflections (a generally positive finding for health professions education; Mann et al., 2009), neither content analysis nor predictive modeling indicates *whose* perspectives are being considered. Other work on the same corpus has shown that many students showed a shift of Perspective over time, from considering obligations to their community to discussing responsibility to their patients (Wise, Reza, & Han, 2020). The different configurations in which reflection elements occur indicates a need for personalized learning support to help students move towards deeper reflection; for example, some students will need support in learning to properly analyze an experience before evaluating it, while others may require help connecting with their feelings. Again, use of data science models can be helpful in identifying what kind of support to offer to which students.

Implications for conceptualizing reflection quality: similarities and distinctions across elements

The classifiers built for Description, Analysis, and Evaluation all depended heavily on the use of past-oriented words; Description and Evaluation were also strongly predicted by expression of personal voice, first-person pronouns, and future-oriented words. These findings align with features suggested conceptually by Cui et al. (2019) and found empirically by Kovanovic et al. (2018). In addition to similar patterns of feature importance, these three elements had strong associations with each other, though Description and Evaluation were more closely related and occurred more frequently than Analysis. This suggests the possibility of certain conceptual connections among the three elements.

In contrast to Description, Analysis, and Evaluation, Feeling and Perspective each showed distinct characteristics regarding association with other elements and features importance. The Feeling classifier depended on some similar features to Description, Analysis, and Evaluation (past-oriented words, text length, expression of personal voice); however, it was also predicted by affective features suggested conceptually by Cui et al. (2019) such as words related to positive emotions and anxiety. The Perspective classifier was quite different, being highly predicted by the use of third-person pronouns (also aligning with Cui et al., 2019). In addition, Perspective was also predicted by expression of personal voice, positive emotion, time, and work, indicating a unique linguistic profile that implies conceptual distinction from other elements. These findings contribute to a better understanding of the nature of reflection and illustrate the value of theory-*informed* data science methods (Tolsgaard et al., 2020) to also be theory-*informing* (Wise & Cui, 2018). The Outcome element was pervasive across reflections. While this was addressed for model-building purposes with resampling, overfitting still occurred as seen in the dramatic decline in the trained model performance on the test data. Thus, the role and importance of assessing Outcome as an element of reflection quality in the current context need to be reconsidered. It may be that, due to its inclusion in almost all reflections, its assessment and support is not a priority.

Multi-label classification was adopted as an approach to improve modeling by taking association among elements into account (Herrera et al., 2016). However, the lack of performance improvement indicates that issues with association cannot be addressed by modeling element covariance. Rather there is a need to clarify the conceptual and empirical distinctions between them. Possibilities for dimensionality reduction (for example collapsing Description and Evaluation, eliminating Outcome) can be explored through principal component analysis. However, such work should be approached carefully since the current data may be limited by important theoretical differences not yet captured empirically. Thus, it may be that such conceptual distinctions between elements and their operationalization for coding first need to be refined. If better element distinction can be achieved in manual coding, then it may be possible to identify additional linguistic features to distinguish them; for example, part-of-speech tagging or coherence patterns which consider the roles of words in sentences and their syntactic relationships (Gibson et al., 2016; Kovanovic et al., 2018).

Implications for conceptualizing reflection quality: what does reflection "depth" represent?

This study used Mezirow's (1991) conceptualization of reflection depth, as operationalized by Tsingos et al. (2015), in which the highest level offered the richest explanation of experiences and their implications. This differs from recent work by Kovanovic et al. (2018) which drew on Hulsman et al. (2009) conceptualization in which the highest level of depth was related to specifying the goals. This may explain why in the current study, the classifier for depth built from linguistic features was strongly driven by the presence of past-oriented words while Kovanovic et al. (2018) reported that past-oriented words predicted the middle, but not highest level of reflection depth. Using Kovanovic et al. scheme, the high presence of Outcome in the current study would have led to misleadingly high levels of reflection depth. This highlights the critical importance of making and communicating conceptual decisions underlying data science approaches (Wise & Cui, 2018).

Relating overall depth to the six reflection elements, the elements-based classifier depended heavily on Description, serving almost as a binary divider between No and Shallow/Deep reflection. Analysis and Evaluation were more associated with Deep than Shallow reflections, aligning both with conceptual notions of depth (Mezirow, 1991) and the empirical results of Kovanovic et al. (2018) who found the highest level of depth was more associated with causal and perceptual words. The overall association of depth with Description, Analysis and Evaluation but not Perspective or Feeling suggests that global reflection quality may currently be evaluated based more on cognitively oriented elements than affective ones.

This leads us to ask a critical question: *What does (and should) reflection depth represent?* Specifically, there is a disconnect between what is asserted as important for reflection theoretically (six elements, see Cui et al., 2019) and what is actually considered when assessing reflection quality in a global way as overall "depth." Similar disconnects between high-level assessment and detailed criteria have been found in other contexts (Ochoa & Duval, 2009; Ochoa et al., 2018). In light of these findings, some composite of reflection elements could offer a better indicator of overall reflection quality than holistic assessments of depth.

Implications for supporting personalized learning for reflective writing

In addition to the contributions to conceptualizing reflection quality and understanding its development over time, the classification models built in this study offer a starting point for developing personalized learning support. We thus now turn our attention to mechanisms by which the application of such models could help inform, support, and improve student reflection.

One straightforward approach would be to initially use the models to detect whether students include no, shallow, or deep reflection in their writing and provide this information to them as a basic overall assessment. Then, the reflection elements could be assessed to provide more detailed information about the aspects of reflection students to which students need to attend. How such information is best aggregated and presented is an important area for future research. For example, the results of this study suggested a tiered structure for feedback where there might be an initial check for the presence of Description; and if it is not present then feedback could focus solely on this foundational element. If Description is found, feedback could progress to Analysis/Evaluation (making sure the latter is grounded in the former), and then Perspectives and Feeling. The choice of elements to emphasize could also be tailored to specific kinds of reflective prompts, and all prompts could include explicit guidance about the elements to include.

Students could also be empowered to use the analytics more actively by filtering their reflections based on element presence, allowing them to explore their own reflective patterns and progress. The system could allow students to save such patterns, annotate them with their comments and create an action plan for what they want to improve. To complete the loop, a system might also remind students of the action plans they set and later invite them to examine if they have achieved their goals. Engaging students as co-designers of such a tool is an important step to ensure the creation of a system that students find both useful and usable (Buckingham Shum et al., 2019).

Finally, in addition to supporting reflection-*on*-(reflective)-practice *after* students write their reflections, there are possibilities to support reflection-*in*-(reflective)-practice *while* students write them. For example, students can be invited to self-assess their reflections for the different elements before submitting and provided with the model's assessment for comparison. Importantly, students should have the ability to indicate when (and why) they disagree with model results, balancing power between the human and technological partners in the system. Discrepancies in these judgements can also be used to improve model performance (Gibson et al., 2016).

Limitations and future work

Limitations of the current study relate to the nature of reflective text as data, the specificities of the reflection context, methodological choices made in cleaning and analyzing the data, and overall questions of generalizability.

First, this study examined student reflections as data representing students' authentic perceptions and thoughts on their professional and academic development. However, the content of what students wrote in their reflection may also relate to what they thought instructors wanted to hear from them (Cotton, 2001) or be shaped by how they wanted to be seen by other students (except in the case of the two "private" reflections). In addition, the wording of the reflective prompts can (intentionally or otherwise) strongly influence what students think and write about (Davis, 2000). The prompts used here introduced notions of professional progress and ethics that may have elicited particular kinds of comments from students. Different kinds of reflective prompts (for example asking students to take others' perspective or include their feelings) may lead students to show different profiles of reflective element presence and overall reflective depth in reflection (though as the prompts remained constant across years, this would not account for the changes observed over time in this study). Finally, as noted above, not all students completed all reflections in all time periods. To the extent such missing data is not random, it is important to interrogate what subpopulations of students are not represented in the data, and thus what ideas and perspectives may be missing.

Second, in working with the reflections, it was necessary to make several decisions for cleaning and analysis that may have impacted the findings. First, the original data corpus included a sizable proportion of reflections that were duplicates of each other, often for corresponding public/private prompts but also in some cases for the same prompt over time. These were treated as non-valid data and deleted prior to analysis based on the premise that students copy-pasted as an expedient approach to a required task; however it is possible that students actually felt the same way at multiple points in time. In addition, care was taken to sample reflections representatively across prompts and time; however the lack of reflections containing No Outcome necessitated the use of oversampling methods to address class imbalance. This may explain the problems of overfitting found for this model and suggests a need to constrict the coding criteria for this element in the content analysis scheme and/or revise the wording of the reflective prompts. Finally, while taking the entire reflection as the unit of analysis was useful to detect the presence of elements and depth expressed across different sentences (Moon, 2013), it limits the specificity with which feedback could be provided to students on their reflections. Future efforts can explore the added value of models with more fine-grained units of analysis (such as the sentence) to inform students about which part of their reflections shows the presence or absence of a particular element and needs to be improved (Knight et al., 2018).

With respect to generalizability, testing the models built over time revealed reduced model performance for several elements in the first two years of the program. This can be explained by the combination of changing language use over time with a model more driven by the (larger) number of reflections made in the later years. For example, in the case of Perspective, the poor performance in the first year may be explained by the differences over time in focusing on community perspectives versus patient ones (Wise et al., 2020). This issue can be addressed either by training models on a corpus with oversampling from the early years or building separate classifiers for early or late program time periods.

In addition, it remains to be investigated the extent to which the specific patterns of reflection development observed here hold for dental students in other cohorts, at other institutions, and to different health professions education contexts more broadly. Thus, a first step for future work is to evaluate model performance across different learning contexts and make refinements where needed. In situations where model performance is suboptimal (or conceptualizations of reflection differ), the theory-informed process documented here can be used to create new models appropriate to these contexts.

Conclusion

Reflection is a critical skill in health professions education to help students become thoughtful practitioners; yet reflection is rarely meaningfully assessed and students are seldom given feedback to develop reflective skills. Working towards an empirically-informed conceptualization of student reflection with the goal of eventually offering personalized support, this study makes several contributions to the growing knowledge base about student reflection in health professions education. First, it empirically established a relationship between overall reflection quality (Depth) and several cognitively oriented elements (mostly Description, followed by Analysis and Evaluation). Second, it probed how the presence of elements and overall reflective depth changed over the course of four academic years, documenting a sharp rise in Shallow reflection (associated with a rise in Description) at the end of the first year, but smaller gains in Deep Reflection, Feeling and Perspective. Additionally, it took critical steps towards personalized reflection support by developing machine learning models that offer reliable discrimination of Shallow versus No Reflection (and the presence of Description and Evaluation). Detection of Deep Reflection and the presence of Analysis, Feelings and Perspectives can be further improved. Together, these efforts further the larger pursuit of helping health professions students become reflective practitioners and lifelong learners.

Appendix

Appendix A: model performance of the classifiers

See Tables 3, 4, 5, 6, 7 and 8.

Table 3 Model performance of the reflection elements using linguistic features

Element	Data	Presence rate (%)	Accuracy	Kappa	Recall	Precision
Description	Train	70	0.85	0.63	0.91	0.88
	Test	68	0.87	0.70	0.94	0.88
Analysis	Train	46	0.76	0.51	0.70	0.75
	Test	43	0.76	0.50	0.67	0.74
Feeling	Train	37	0.78	0.50	0.59	0.76
	Test	36	0.80	0.54	0.60	0.79
Perspective	Train	64	0.75	0.44	0.86	0.78
	Test	66	0.79	0.53	0.85	0.83
Evaluation	Train	58	0.80	0.60	0.83	0.83
	Test	56	0.83	0.65	0.86	0.84
Outcome	Train	66*	0.96	0.90	0.97	0.96
	Test	63*	0.80	0.53	0.93	0.79

*After class rebalancing using SMOTE

Table 4 Model performance of the depth classifier using linguistic features

Data	Presence rate (deep/shallow/no)	Accuracy	Kappa	Recall*	Precision*
Train	0.17/0.53/0.30	0.74	0.55	0.71	0.73
Test	0.14/0.52/0.34	0.79	0.63	0.79	0.80

*Calculation was made by taking a weighted average of each class's precision/recall by the number of cases of each class

Table 5 Confusion matrix of the depth classifier using linguistic features on the test data

Actual	Predicted			Recall	Precision
	Deep reflection	Shallow reflection	No reflection		
Deep reflection	31	6	0	0.72	0.84
Shallow reflection	12	133	29	0.85	0.76
No reflection	0	17	72	0.71	0.81

Table 6 Model performance of the depth classifier using coded and predicted values for reflection elements as features

Data	Presence rate (deep/shallow/no)	Accuracy	Kappa	Recall*	Precision*
Train	0.17/0.53/0.30	0.83	0.72	0.76	0.76
Test coded	0.14/0.52/0.34	0.82	0.70	0.72	0.71
Test predicted		0.73	0.53	0.69	0.67

*Calculation was made by taking a weighted average of each class's precision/recall by the number of cases of each class

Table 7 Confusion matrix of the depth classifier using coded values for reflection elements on the test data

Actual	Predicted			Recall	Precision
	Deep reflection	Shallow reflection	No reflection		
Deep reflection	17	21	0	0.40	0.45
Shallow reflection	26	135	6	0.87	0.81
No reflection	0	0	95	0.94	1.00

Table 8 Confusion matrix of the depth classifier using predicted values for reflection elements on the test data

Actual	Predicted			Recall	Precision
	Deep reflection	Shallow reflection	No reflection		
Deep reflection	21	18	0	0.49	0.54
Shallow reflection	22	127	31	0.81	0.71
No reflection	0	11	70	0.69	0.87

Appendix B: lists of top 10 predictive features for each of the classifiers

See Tables 9, 10 and 11.

Table 9 Reflection elements classifiers: Top 10 features (with MDG mean index)

	Description		Analysis		Feeling		Perspective		Evaluation		Outcome		Multi-label classifier	
# 1	focuspast	105.17	focuspast	90.78	focuspast	51.36	they	48.49	focuspast	118.94	i	36.73	authentic	0.27
# 2	authentic	41.44	wc	51.38	wc	37.35	authentic	31.54	authentic	53.42	clout	7.71	focuspast	0.23
# 3	i	24.51	authentic	27.49	authentic	26.55	posemo	22.96	wc	23.78	affiliation	7.05	i	0.22
# 4	focusfuture	23.33	discrep	21.61	feel	18.98	time	20.83	discrep	20.26	we	6.66	time	0.22
# 5	time	17.49	time	16.18	reward	17.97	work	19.31	i	20.01	article	5.92	WC	0.17
# 6	wc	17.21	quant	14.51	anx	15.83	wc	19.11	focusfuture	18.20	authentic	4.68	clout	0.17
# 7	discrep	15.01	cause	13.28	clout	15.27	dic	16.53	clout	17.24	health	3.50	feel	0.15
# 8	quant	13.00	dic	10.96	posemo	14.33	clout	16.38	time	13.61	focusfuture	3.33	discrep	0.15
# 9	prep	10.99	insight	10.77	i	14.21	tone	15.69	quant	13.43	risk	3.27	quant	0.15
# 10	clout	8.76	i	10.61	time	11.18	quant	12.15	space	10.50	you	3.09	analytic	0.11
Mean (SD)	4.5 (12.4)		5.7 (10.8)		5.2 (6.3)		5.1 (6.1)		5.4 (14.0)		1.4 (4.3)		0.038 (0.065)	

Table 10 Depth classifier using linguistic feature: Top 10 features (with MDG mean index)

	Feature	MDG	Mean feature values (SD)		
			Deep reflection	Shallow reflection	No reflection
# 1	focuspast	102.12	3.3 (2.1)	2.3 (2.1)	0.5 (0.9)
# 2	WC	72.85	172.4 (55.1)	105.6 (46.5)	85.2 (40.5)
# 3	Authentic	41.32	75.5 (22.9)	70.9 (27.6)	42.9 (29.5)
# 4	focusfuture	24.31	1.4 (1)	1.6 (1.5)	2.4 (2.2)
# 5	i	22.15	8.5 (3.3)	8.9 (3.8)	6.4 (4.5)
# 6	discrep	19.94	1.4 (1.2)	1.6 (1.6)	2.3 (1.9)
# 7	time	16.81	5.4 (2.4)	5.1 (3.2)	3 (2.3)
# 8	focuspresent	14.62	11.4 (3.2)	12.7 (3.8)	11.6 (4)
# 9	prep	13.61	15.5 (2.6)	15.6 (3.2)	16.5 (3.9)
# 10	quant	13.21	2.6 (1.6)	2.7 (2)	1.7 (1.7)

Table 11 Depth classifier using coded reflection elements: Top 10 features (with MDG mean index)

	Feature	MDG	Mean (SD)		
			Deep reflection	Shallow reflection	No reflection
# 1	Description	400.78	1.00 (0)	1.00 (0.04)	0.03 (0.17)
# 2	Analysis	29.41	0.92 (0.27)	0.57 (0.50)	0.01 (0.10)
# 3	Evaluation	11.39	0.72 (0.45)	0.45 (0.50)	0.04 (0.19)
# 4	Feeling	11.38	0.72 (0.45)	0.57 (0.50)	0.73 (0.44)
# 5	Perspective	9.65	0.98 (0.16)	0.77 (0.42)	0.02 (0.13)
# 6	Outcome	1.95	1.00 (0)	0.97 (0.18)	0.84 (0.37)

Acknowledgements The authors would like to thank Sarita Xinying Liu for her assistance with the process of content analysis and the NYU Dental Informatics team for their help in accessing the data. This article expands on preliminary work presented at the International Conference on Learning Analytics and Knowledge in 2020.

Declarations

Conflict of interest None.

References

Alzahem, A. M., Van der Molen, H. T., Alaujan, A. H., Schmidt, H. G., & Zamakhshary, M. H. (2011). Stress amongst dental students: A systematic review. *European Journal of Dental Education, 15*(1), 8–18.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596.

Asadoorian, J., Schönwetter, D. J., & Lavigne, S. E. (2011). Developing reflective health care practitioners: Learning from experience in dental hygiene education. *Journal of Dental Education, 75*(4), 472–484.

Boud, D., Keogh, R., & Walker, D. (Eds.). (1985). *Reflection: Turning experience into learning*. Kogan Page.
Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32.
Buckingham Shum, S., Ferguson, R., & Martinez-Maldonado, R. (2019). Human-centred learning analytics. *Journal of Learning Analytics, 6*(2), 1–9.
Bush, H., & Bissell, V. (2008). The evaluation of an approach to reflective learning in the undergraduate dental curriculum. *European Journal of Dental Education, 12*(2), 103–110.
Cantor, A. B. (1996). Sample-size calculations for cohen's kappa. *Psychological Methods, 1*(2), 150–153.
Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.
Chirema, K. D. (2007). The use of reflective journals in the promotion of reflection and learning in post-registration nursing students. *Nurse Education Today, 27*(3), 192–202.
Cotton, A. H. (2001). Private thoughts in public spheres: Issues in reflection and reflective practices in nursing. *Journal of Advanced Nursing, 36*(4), 512–519.
Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods, 49*(3), 803–821.
Cui, Y., Wise, A. F., & Allen, K. L. (2019). Developing reflection analytics for health professions education: A multi-dimensional framework to align critical concepts with data features. *Computers in Human Behavior, 100*, 305–324.
Davis, E. A. (2000). Scaffolding students' knowledge integration: Prompts for reflection in KIE. *International Journal of Science Education, 22*(8), 819–837.
Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process*. D.C. Heath.
Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends, 59*(1), 64–71.
Gibbs, G. (1988). *Learning by doing: A guide to teaching and learning methods*. Oxford, UK.
Gibson, A., Aitken, A., Sándor, Á., Buckingham Shum, S., Tsingos-Lucas, C., & Knight, S. (2017). Reflective writing analytics for actionable feedback. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 153–162). New York, NY: ACM.
Gibson, A., Kitto, K., & Bruza, P. (2016). Towards the discovery of learner metacognition from reflective writing. *Journal of Learning Analytics, 3*(2), 22–36.
Hanson, K., & Alexander, S. (2010). The influence of technology on reflective learning in dental hygiene education. *Journal of Dental Education, 74*(6), 644–653.
Heeneman, S., & Driessen, E. W. (2017). The use of a portfolio in postgraduate medical education–reflect, assess and account, one for each or all in one? *GMS Journal for Medical Education, 34*(5), 1–12.
Herrera, F., Charte, F., Rivera, A. J., & Del Jesus, M. J. (2016). *Multilabel classification*. Springer.
Hulsman, R. L., Harmsen, A. B., & Fabriek, M. (2009). Reflective teaching of medical communication skills with DiViDU: Assessing the level of student reflection on recorded consultations with simulated patients. *Patient Education and Counseling, 74*(2), 142–149.
Ip, W. Y., Lui, M. H., Chien, W. T., Lee, I. F., Lam, L. W., & Lee, D. (2012). Promoting self- reflection in clinical practice among Chinese nursing undergraduates in Hong Kong. *Contemporary Nurse*, *41*(2), 253–262.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.
Knight, S., Shum, S. B., Ryan, P., Sándor, Á., & Wang, X. (2018). Designing academic writing analytics for civil law student self-assessment. *International Journal of Artificial Intelligence in Education, 28*(1), 1–28.
Koole, S., Dornan, T., Aper, L., Scherpbier, A., Valcke, M., Cohen-Schotanus, J., & Derese, A. (2011). Factors confounding the assessment of reflection: A critical review. *BMC Medical Education, 11*(1), 104.
Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G., and Dawson, S. (2018). Understand students' self-reflections through learning analytics. In *Proceedings of the 8th international conference on learning analytics & knowledge* (pp. 389–398). New York, NY, USA: ACM.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.
Liu, M., Shum, S. B., Mantzourani, E., & Lucas, C. (2019). Evaluating Machine Learning Approaches to Classify Pharmacy Students' Reflective Statements. In Proceedings of *International Conference on Artificial Intelligence in Education* (pp. 220–230). Cham, Switzerland: Springer.
Mann, K., Gordon, J., & MacLeod, A. (2009). Reflection and reflective practice in health professions education: A systematic review. *Advances in Health Sciences Education, 14*(4), 595.
Mezirow, J. (1991). *Transformative dimensions of adult learning*. Jossey- Bass.

Moon, J. A. (2013). *Reflection in learning and professional development: Theory and practice*. Routledge.
Ochoa, X., Domínguez, F., Guamán, B., Maya, R., Falcones, G., & Castells, J. (2018). The rap system: automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors. In *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 360–364). New York, NY, USA: ACM.
Ochoa, X., & Duval, E. (2009). Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries, 10*(2–3), 67–91.
Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin.
Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning, 85*(3), 333.
Schön, D. A. (1983). *The reflective practitioner*. Jossey-Bass.
Shibani, A., Knight, S., & Shum, S. B. (2019). Contextualizable learning analytics design: A generic model and writing analytics evaluations. In *Proceedings of the 9th international conference on learning analytics and knowledge* (pp. 210–219). New York, NY, USA: ACM.
Tolsgaard, M. G., Boscardin, C. K., Park, Y. S., Cuddy, M. M., & Sebok-Syer, S. S. (2020). The role of data science and machine learning in Health Professions Education: practical applications, theoretical contributions, and epistemic beliefs. *Advances in Health Sciences Education*. https://doi.org/10.1007/s10459-020-10009-8
Tsingos, C., Bosnic-Anticevich, S., Lonie, J. M., & Smith, L. (2015). A model for assessing reflective practices in pharmacy education. *American Journal of Pharmaceutical Education, 79*(8), 124.
Ullmann, T. D. (2019). Automated analysis of reflection in writing: validating machine learning approaches. *International Journal of Artificial Intelligence in Education, 29*(2), 217–257.
Wise, A., & Cui, Y. (2018). Envisioning a learning analytics for the learning sciences. In *Proceedings of the 13th international conference of the learning sciences* (pp. 1799–1806). London, UK: International Society of the Learning Sciences.
Wise, A.F., Reza, S. & Han, R. J. (2020). Becoming a dentist: Tracing professional identity development through mixed-methods data mining of student reflections. In *Proceedings of the 13th international conference of the learning sciences*. Nashville, TN: International Society of the Learning Sciences.
Wise, A. F., & Shaffer, D. W. (2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics, 2*(2), 5–13.
Wong, F. K., Kember, D., Chung, L. Y., & Yan, L. (1995). Assessing the level of student reflection from reflective journals. *Journal of Advanced Nursing, 22*(1), 48–57.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.