Automated Analysis of Middle School Students' Written Reflections During Game-Based Learning

Dan Carpenter[1], Michael Geden[1], Jonathan Rowe[1], Roger Azevedo[2], and James Lester[1]

[1] North Carolina State University, Raleigh, NC 27695, USA
{dcarpen2, mageden, jprowe, lester}@ncsu.edu

[2] University of Central Florida, Orlando, FL 32816, USA
roger.azevedo@ucf.edu

Abstract. Game-based learning environments enable students to engage in authentic, inquiry-based learning. Reflective thinking serves a critical role in inquiry-based learning by encouraging students to think critically about their knowledge and experiences in order to foster deeper learning processes. Free-response reflection prompts can be embedded in game-based learning environments to encourage students to engage in reflection and externalize their reflection processes, but automatically assessing student reflection presents significant challenges. In this paper, we present a framework for automatically assessing students' written reflection responses during inquiry-based learning in CRYSTAL ISLAND, a game-based learning environment for middle school microbiology. Using data from a classroom study involving 153 middle school students, we compare the effectiveness of several computational representations of students' natural language responses to reflection prompts—GloVe, ELMo, tf-idf, unigrams—across several machine learning-based regression techniques (i.e., random forest, support vector machine, multi-layer perceptron) to assess the depth of student reflection responses. Results demonstrate that assessment models based on ELMo deep contextualized word representations yield more accurate predictions of students' written reflection depth than competing techniques. These findings point toward the potential of leveraging automated assessment of student reflection to inform real-time adaptive support for inquiry-based learning in game-based learning environments.

Keywords: Reflection, Self-regulated Learning, Metacognition, Game-Based Learning, Natural Language.

1 Introduction

Game-based learning environments provide rich opportunities for students to engage in scientific inquiry by exploring problems that are complex, open-ended, and realistic [1]. Inquiry-based learning has been demonstrated to yield significant improvements in students' science literacy and research skills [2, 3]. However, the open-ended nature of inquiry learning in game-based environments can prove challenging for many students, which points toward the importance of students effectively regulating their own learning processes [4-6]. Reflection is a key component of self-regulated learning [7]. During reflection, students can become aware of their problem-solving progress and make adaptations to their learning strategies, which can lead to improved learning outcomes [8-10]. We define reflection as a process of introspective consideration of one's own knowledge and learning experiences, which is used to inform strategic revisions for improving learning [11]. During inquiry-based learning, it is important for students to reflect on their knowledge and actions to ensure that they are on track to achieving their desired learning objectives.

A common approach to capturing students' reflections during learning is through free-response reflection prompts [12]. Free-response reflection prompts can be embedded in game-based learning environments to encourage reflection and externalize students' reflection processes. A key dimension of student reflection is reflection depth, which distinguishes between responses that exemplify productive reflection versus surface-level observations or verbatim restatements of content [13, 14].

Assessing students' written responses to reflection prompts can provide insight into the characteristics of students' reflective thinking. However, assessing students' written reflections is often a manual, labor-intensive process. Devising automated methods for assessing reflection is critical for enabling adaptive learning environments that can support students' self-regulatory processes during inquiry-based learning. Approaches to automatically assessing student reflection include expert-crafted rule-based systems, dictionary-based techniques that search for specific words and phrases, and machine learning approaches that are data-driven [15]. Machine learning approaches show particular promise for devising automated reflection assessment models that are accurate, reliable, and can be utilized at run-time [15]. Previous work investigating machine learning approaches to automatically assessing written reflections has used count-based representations of students' natural language reflections [15, 16]. Recent advances in distributed embedding-based representations of natural language show particular promise for encoding students' natural language reflections for automated assessment [17, 18]. Using pre-trained word embeddings, such as GloVe [19] and ELMo [20], syntactic and semantic information captured from large corpora can be leveraged to concisely represent students' written reflections.

In this paper, we present a framework for automatically assessing students' written reflections during inquiry-based learning. Using written reflections of 153 middle school students, we investigate several vector-based representations of students' written reflection responses—unigram, tf-idf, GloVe, and ELMo embedding-based representations—to induce machine learning-based models for measuring the depth of student reflection.

2 Related Work

Reflection plays a critical role in self-regulated learning (SRL). In the Information Processing Theory of SRL [7], reflection is both a backward-looking and forward-looking process [21]. Specifically, students look back at what they have learned and the actions they have taken in the past, and they consider what changes they might need to make to achieve their learning goals moving forward [21]. Reflection is especially important in inquiry-based learning, since it is important for students to understand the relationships between their learning and problem-solving goals [22].

A common approach for assessing students' written reflections is to create a model that distinguishes between varying degrees of reflection depth and different characteristics of reflection breadth [12]. In surveying 34 different models used to analyze reflection, Ullmann [12] found that many models include some notion of reflective depth, often ranging from non-reflective to slightly reflective to highly reflective [13, 23]. Many models also attempt to capture the breadth of reflection, including aspects such as 'attending to feelings' and 'validation' [24] or 'justification' [25]. Students' written responses to reflection prompts embedded in game-based learning environments are often brief, and therefore, inherently limited in reflective breadth. Thus, reflective depth serves as a proxy for measuring the quality of students' reflective thinking during inquiry-based learning in game-based environments.

After establishing a model of reflection, a manual coding process is commonly used to analyze and assess written reflections [15]. Coding students' written reflections can be a labor-intensive process, which has motivated growing interest in automated reflection analysis methods. Approaches to automatic reflection assessment include dictionary-based, rule-based, and machine learning-based systems [15, 16]. Prior work on automated analysis of student reflections has largely used one-hot encodings and features derived from LIWC and Coh-Metrix to represent students' reflections [15, 16]. However, recent advances in natural language understanding and automated essay scoring suggest that pre-trained word embeddings, such as GloVe [19] and ELMo [20], show promise as representations of students' written reflections [17, 18], since they are trained on large corpora and capture both syntactic and semantic aspects of language. Of the work that has been done to automatically assess written reflection, there is a common focus on assessing the written reflections of students in higher education [15, 16]. While supporting reflection in college students is important, substantial benefits can be found when students engage in SRL processes from a young age [26, 27]. Written data from K-12 students presents a distinctive set of challenges, since it is often short and rife with grammatical errors and misspellings [28]. There is a need to investigate more robust techniques for representing written reflections of K-12 students.

Two recent studies, conducted by Kovanovic et al. [16] and Ullmann [15], have investigated machine learning-based approaches for automated assessment of student reflections. Kovanovic et al. [16] coded three different types of reflections (i.e., observations, motives, and goals). To represent written reflections, they extracted the 100 most common unigrams, bigrams, and trigrams (300 total) from their corpus, generated linguistic features using the Linguistic Inquiry and Word Count (LIWC) tool, and extracted several Coh-Metrix features [16]. The model of reflection used by Ullmann [15] included a binary indicator of reflective depth (i.e., reflective versus non-reflective) and seven breadth dimensions that address common components of reflective models (e.g., description of an experience, awareness of difficulties, and future intentions). Ullmann used binary vectors to represent the unique unigrams that occurred in each reflection, ignoring any unigrams that occurred less than ten times throughout the entire corpus [15].

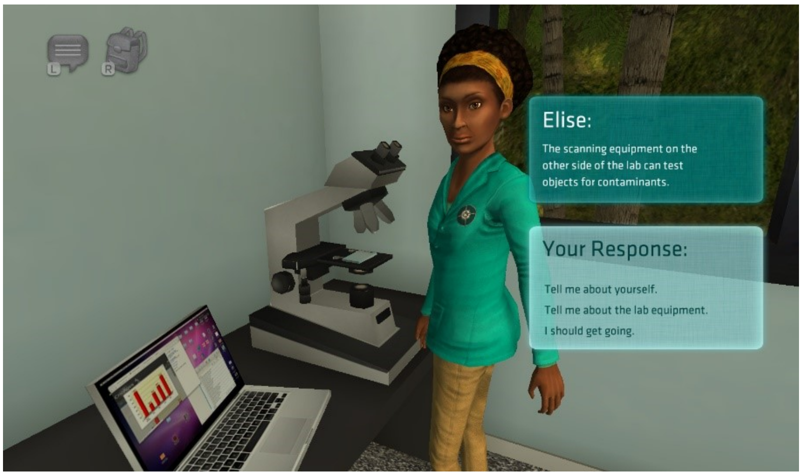
In contrast to this previous work, our model of reflection evaluates reflection depth on a continuous scale. We use Ullmann's binary unigram representation of written reflection as a baseline and investigate the benefits of several language modeling techniques: tf-idf, GloVe, and ELMo. Tf-idf represents a step up in complexity from the binary unigram representation and has been used as a baseline representation for text classification [29]. GloVe [19] and ELMo [20] concisely capture both syntactic and semantic aspects of language. For GloVe and ELMo, we represent student reflections as the average of the embeddings for each word [30]. Furthermore, Kovanovic et al. [16] and Ullmann [15] investigated written reflections collected from undergraduate students, while we explore middle school students' reflections as they engage with a game-based learning environment in their science classrooms.

3 Method

To investigate automated assessment of student reflection, we use data from student interactions with CRYSTAL ISLAND, a game-based learning environment for middle school microbiology (Fig. 1). In CRYSTAL ISLAND, students adopt the role of a science detective who has recently arrived at a remote island research station to investigate the cause of an outbreak among a group of scientists. Students explore the open-world virtual environment, gather information by reading in-game books and articles, speak with non-player characters, perform scientific tests in a virtual laboratory, and record their findings in a virtual diagnosis worksheet. Students solve the mystery by submitting a diagnosis explaining the type of pathogen causing the illness, the transmission source of the disease, and a recommended treatment or prevention plan.

3.1 Student Reflection Dataset

We analyze a corpus of students' written reflections collected during a pair of classroom studies involving middle school students interacting with CRYSTAL ISLAND



Elise:

The scanning equipment on the other side of the lab can test objects for contaminants.

Your Response:

Tell me about yourself.

Tell me about the lab equipment.

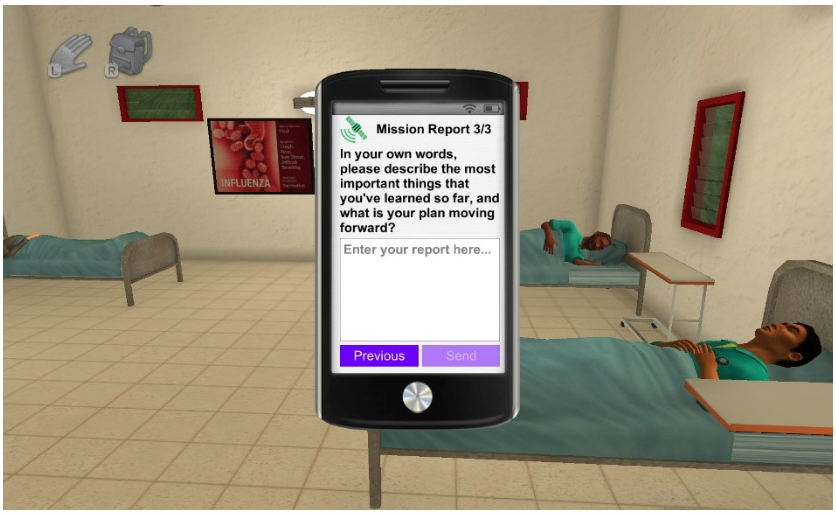
I should get going.

Fig. 1. Crystal Island game-based learning environment.

during spring 2018 and spring 2019. Data was collected from 153 students in total, but only 118 students reported demographic information. Among these students, 51% identified as female, and ages ranged from 13-14 (M=13.6, SD=0.51). 43 students reported being Caucasian/White, 32 reported being African American, 21 students reported being Hispanic or Latino, and 3 reported being of Asian descent. The students did not have prior experience with Crystal Island.

In both studies, students completed a series of pre-study measures the week prior to interacting with Crystal Island, including a microbiology content knowledge test, an emotions, interest, and value instrument, and an achievement goal instrument. Students were briefly introduced to the game by a researcher, and they viewed a short video trailer that provided background on the game's storyline. Afterward, students interacted with Crystal Island until they solved the mystery or when approximately 100 minutes of gameplay time had elapsed. After finishing the game, students completed a series of post-study materials, which included another microbiology content knowledge test as well as several questionnaires about students' experiences with the game, including sense of presence and engagement.

While interacting with Crystal Island, students were periodically prompted to reflect on what they had learned thus far and what they planned to do moving forward (Fig. 2). These reflection prompts came after major game events, such as talking with the camp nurse, testing objects in the virtual laboratory, or submitting a diagnosis. Students received several prompts for reflection during the game (M=3.0, SD=0.95). After completing the game or running out of time, students were asked to reflect on their problem-solving experience as a whole, explaining how they approached the problem and whether they would do anything differently if they were asked to solve a similar problem in the future. In total, the data included 728 reflection responses from 153 students. The average length of a reflection response was approximately 19 words (min=1, max=100, SD=14.2). (Please see Table 1 for several example student responses to reflection prompts in Crystal Island.)



Mission Report 3/3

In your own words, please describe the most important things that you've learned so far, and what is your plan moving forward?

Enter your report here...

Fig. 2. In-game reflection prompt presented to students.

3.2 Annotating Students' Written Responses to Reflection Prompts

To measure the depth of students' responses to reflection prompts, a five-point scale was developed by two of the authors using a grounded theory approach [31]. The scale was devised to measure the extent to which students assessed their own knowledge and articulated plans exemplifying high-quality reasoning, hypothesis formation, and metacognition. The researchers reviewed 20 reflection responses together, discussing the strengths and weaknesses of each. These reflection responses were individually selected to represent the range of reflection depth in the dataset, with the goal of including several reflections for each of the five ratings. That is, the researchers selected some reflections that seemed to be particularly weak and discussed why they were weak. The observations and insights from these discussions formed the basis for the lowest reflection depth rating. A similar process was used for the other ratings to develop a rubric for evaluating reflection depth (Table 1), providing examples and reasoning for the five possible scores. Once the rubric was developed, the researchers separately annotated another 20 reflections to verify the reliability of the model, then discussed and reconciled incongruent ratings. Finally, the remaining 708 reflections were separately annotated by both researchers and an intraclass correlation of 0.669 was achieved, indicating moderate inter-rater reliability. The final ratings of reflection depth were calculated by averaging the values assigned by the two authors (M=2.41, SD=0.86), yielding a continuous measure of reflection. Averaging ratings is a standard approach for reconciling differences between coders' assigned ratings, although it does have limitations. For example, reflections that received the same rating from both coders (e.g., 3 and 3) and reflections that received different ratings (e.g., 2 and 4) would be rated the same even though there is disagreement in the latter case.

3.3 Modeling Reflective Depth using Natural Language Embeddings

Prior to modeling student reflections, the text responses were normalized using tokenization, conversion to lowercase, and removal of non-grammatical characters. When generating binary unigram vectors, tokens that appeared fewer than ten times throughout the corpus were removed. Similarly, any words that were not found in the GloVe embeddings were ignored when calculating average GloVe and ELMo word embeddings, effectively removing misspelled words from the data. We trained regression models using random forests, SVM, and feedforward neural networks using scikit-learn [32]. Reflection assessment models were trained using nested 10-fold cross-validation at the student level. Within each fold, 10-fold cross-validation was used for hyperparameter tuning. Random forest models were tuned over the number of trees in the forest (100, 200, or 300), the minimum number of samples required to split an internal node (2, 4, or 10), and a the maximum number of features to consider when searching for the best split (log2 or no maximum). SVM models were tuned over the kernel type (rbf or linear) and the regularization parameter (1, 2, 5, 10). Multi-layer perceptron models were tuned over the number of neurons in the hidden layer (50, 100, or 200) and the L2 regularization penalty (0.0001, 0.001, 0.01).

As a baseline, we encoded each natural language reflection as a binary vector representing the unique unigrams that occurred in that reflection (i.e., a one-hot encoding).

Table 1. Rubric used to annotate students' written responses to reflection prompts. Reflections showing at least one characteristic in the middle column were assigned the associated rating.

RATING	CHARACTERISTICS	EXAMPLES
1	Lacks both a plan and knowledge; abstract and largely meaningless; unactionable	"Each clue will help with solving the problem"; "Yeah cool game I learned science"
2	Presents a vague hypothesis or plan with no clear reasoning; simply restates information that was directly learned in the game, with no abstraction or inference on the part of the student	"That the illness causing the people being sick might be pathogen"; "I found out that the egg has bacteria"; "I think I am going to talk to other people"
3	Presents a clear hypothesis or a plan, but doesn't provide any reasoning for it; demonstrates awareness about gaps in knowledge and presents a plan to fix it; organizes the importance of their knowledge	"Getting more information off the food I think it has something to do with the food"; "The most important thing is how the illness is spreading"
4	Presents a clear hypothesis or plan with reasoning; provides an abstraction of the situation with a plan; addresses what they have learned, why it is important, and what they plan to do with this information	"I plan on questioning the cook as they know more about the food and how it could be contaminated with viruses or bacteria"; "I need to learn more about what the sick people do on a day to day schedule"
5	Presents both a clear hypothesis and plan with reasoning; presents a high-quality sequence of abstract plans	"I think that it might have to do with salmonella because when I tested the milk it was positive with pathogenic bacteria. I think that I will test things that can be contaminated"; "I will continue to test the foods the sick people touched or previously ate to see if it's contaminated"

This was a 220-dimension vector, where each index represents the presence of a specific word in the corpus vocabulary after infrequent words were removed. We also encoded the student reflections as tf-idf vectors, which are sparse real-valued vectors that represent documents based on the frequency of each term in the corpus, weighted by the uniqueness of that term in the corpus. Since tf-idf accounts for the frequency of each word, unlike the binary unigram representation, infrequent words were not removed. Finally, we examined two different word embedding techniques, GloVe [19] and ELMo [20]. GloVe embeddings are word-based, so it is possible to use pre-trained GloVe embeddings, which have been trained on other corpora (i.e., Wikipedia and Gigaword), and simply look up embeddings by word. We also investigated the benefits of fine-tuning GloVe embeddings. Fine tuning allows you to take the pre-trained embeddings and infuse domain-specific information from an available corpus. Both the pre-trained and fine-tuned GloVe embeddings were 300-dimension real-valued vectors. ELMo, which was also trained on large corpora but uses character-based methods to represent text, is built with the intention that sentences, and not individual words, are used to create embeddings [20]. To maintain a fair comparison between the various representations of students' written reflections, we first embedded entire written reflection responses with ELMo and then extracted individual word embeddings. This allows the embeddings to capture information related to the specific context in which each word was used. The ELMo word embeddings were 256-dimension real-valued vectors. For both GloVe and ELMo, we represented the reflection text as the average embedding across all words in the reflection.

4 Results

To investigate the relationship between student learning outcomes and depth of reflection during inquiry-based learning in CRYSTAL ISLAND, we utilized Pearson correlation analysis. Average reflection depth ratings for all reflections a student wrote were found to be positively correlated with student post-test scores, r(601)=.29, p<.001).

Next, we compared the accuracy of competing models of reflection depth across five natural language embedding representations and three machine learning-based regression techniques. Models were evaluated using R-squared, mean absolute error, and mean squared error (Table 2).

Table 2. Model results using 10-fold cross-validation. Bold values represent best performance.

	RF			***SVM***			***NN-MLP***		
Text Features	**R^2**	**MSE**	**MAE**	**R^2**	**MSE**	**MAE**	**R^2**	**MSE**	**MAE**
Binary unigram	0.57	0.32	0.42	0.62	0.28	0.41	0.49	0.37	0.46
TF-IDF	0.53	0.34	0.43	0.40	0.43	0.51	0.43	0.51	0.55
GloVe	0.49	0.38	0.49	0.48	0.38	0.47	0.38	0.67	0.61
GloVe fine-tuned	0.49	0.38	0.49	0.52	0.35	0.45	0.35	0.62	0.62
ELMo	0.55	0.33	0.45	**0.64**	**0.26**	**0.40**	0.26	0.39	0.49

Results indicated that SVM models using average ELMo embeddings to represent students' written reflections achieved the highest predictive accuracy (R-squared = 0.64, MSE = 0.26, MAE = 0.40). While we expected the tf-idf representation to yield improved performance relative to the binary unigram representation, the top performing model using tf-idf vectors performed substantially worse (R-squared = 0.53, MSE = 0.34, MAE = 0.43). This may be due to the fact that, while tf-idf accounts for infrequent terms, keeping words with fewer than ten occurrences in the corpus resulted in a very large and sparse feature space. We also expected GloVe word embeddings, which are able to leverage data from large corpora, to outperform both binary unigram and tf-idf, but the GloVe embedding representations of students' written reflections generally performed the worst out of all feature representations (R-squared = 0.49, MSE = 0.38, MAE = 0.49). Fine tuning GloVe embeddings using the CRYSTAL ISLAND reflection dataset appears to help (R-squared = 0.52, MSE = 0.35, MAE = 0.45), but the improvement is marginal. Notably, the accuracy of the SVM+ELMo approach was greater than all competing methods, including the binary unigram baseline representation, but the improvement was relatively small. A possible explanation is that the information captured by ELMo's character-level embeddings and sentence-based contextualization is critical, especially considering the small size of the dataset used in this work. In comparison, GloVe produces word-level embeddings that are not contextualized, which means that GloVe embeddings encode less fine-grained information as well as less context-based information. The performance of unigram models may be explained by the fact that they use only data from students' natural language responses to reflection prompts in CRYSTAL ISLAND, which removes potential noise from external data sources.

To better understand how the competing models distinguished between different levels of depth in students' written reflections, we qualitatively examined several select assessments generated by the SVM+ELMo model, several of which are shown below in Table 3.

Table 3. Predictions of reflection depth (SVM with ELMo features).

Reflection	Predicted Score	Actual Score
"The most important things I've learned are that oranges, raw chicken, and tomato were tested positive for nonpathogenic virus. Eggs were tested positive for pathogenic virus. I believe that salmonellosis is the disease that the infected people on Crystal Island have, but I will have to gather some more information on other diseases."	3.3	4
"The egg has a pathogenic virus in it. Influenza is a virus that is spread through direct contact and the only prevention is vaccination."	3.1	3.5
"The milk is contaminated with pathogenic bacteria. To test other foods sick members may have been in contact with."	3.1	3
"I realized that raw chicken has influenza."	1.4	2
"I've learned a lot and my plan moving forward is in progress."	1.4	1

Examples that were assigned higher depth scores appeared to be longer and contain more terms that relate to the microbiology content (e.g., pathogenic, virus, bacteria) in CRYSTAL ISLAND. This is notable because the ELMo embedding representation should not be sensitive to reflection length; it uses the average word embedding of the reflection response. Reflection responses that were assigned lower scores, on the other hand, are shorter and use fewer terms relevant to the learning scenario's science content. Low-scoring reflections are short, vague, and provide little evidence of deeper reasoning.

5 Conclusion and Future Work

Reflection is critical to learning. Scaffolding student reflection in game-based learning environments shows significant promise for supporting self-regulation and enhancing learning outcomes. By prompting students to engage in written reflection during inquiry-based learning experiences, there is an opportunity to identify when students are not reflecting effectively and scaffold their self-regulated learning processes. This work investigated machine learning-based methods for automatically assessing the depth of student reflection by leveraging natural language embedding-based representations (i.e., GloVe and ELMo) of reflections in a game-based learning environment for middle school microbiology education. Results showed that SVM models using average ELMo embeddings were best able to predict reflection depth compared to competing baseline techniques.

There are several promising directions for future research on automated assessment and support of student reflection during inquiry-based learning. First, investigating methods to address the inherent "noisiness" of middle school students' reflective writings, including misspellings, grammatical errors, non-standard word usage, and other issues of writing quality, shows significant promise, as they are an inevitable feature of K-12 student writing. A related direction is to investigate the relationship between students' English language proficiency and the ratings assigned to their written reflections. Another direction for future work is to investigate alternative machine learning techniques for modeling the depth of student reflections, including deep neural architectures (e.g., recurrent neural networks). Deep recurrent neural networks have been found to be especially effective for capturing sequential patterns in natural language data, and it is possible that they may be well suited for modeling sequential linguistic structures that are more indicative of reflection depth than individual words. Moreover, since deep neural networks can learn abstract representations of data, models of student reflection derived using deep neural networks may be able to generalize to written reflections in different domains. Finally, it will be important to investigate ways in which computational models for automatically assessing student reflection can be used to generate explanations for ratings of reflection depth, which can be provided to learners and teachers to help support the development of reflection and self-regulated learning skills.

Acknowledgements. This research was supported by funding from the National Science Foundation under Grant DRL-1661202. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

1. Plass, J., Mayer, R. E., & Homer, B. (Eds.). Handbook of game-based learning. MIT Press, Cambridge, MA (2020).
2. Gormally, C., Brickman, P., Hallar, B., & Armstrong, N. Effects of inquiry-based learning on students' science literacy skills and confidence. International Journal for the Scholarship of Teaching and Learning, 3(2), n2 (2009).
3. Lazonder, A. W., & Harmsen, R. Meta-analysis of inquiry-based learning: Effects of guidance. Review of Educational Research, 86(3), 681-718 (2016).
4. Belland, B. R., Walker, A. E., Kim, N. J., & Lefler, M. Synthesizing results from empirical research on computer-based scaffolding in STEM education: A meta-analysis. Review of Educational Research, 87(2), 309-344 (2017).
5. Yew, E. H., & Goh, K. Problem-based learning: An overview of its process and impact on learning. Health Professions Education, 2(2), 75-79 (2016).
6. Taub, M., Sawyer, R., Smith, A., Rowe, J., Azevedo, R., & Lester, J. The agency effect: The impact of student agency on learning, emotions, and problem-solving behaviors in a game-based learning environment. Computers & Education, 147, 103781 (2020).
7. Winne, P. H. Cognition and metacognition within self-regulated learning. In: Handbook of self-regulation of learning and performance (pp. 52-64). Routledge (2017).
8. Azevedo, R., Mudrick, N. V., Taub, M., & Bradbury, A. E. Self-regulation in computer-assisted learning systems. In: J. Dunlosky & K. Rawson (Eds.), The Cambridge handbook of cognition and education, pp. 587-618. Cambridge Press, Cambridge, MA (2019).
9. Joksimović, S., Dowell, N., Gašević, D., Mirriahi, N., Dawson, S., & Graesser, A. C. Linguistic characteristics of reflective states in video annotations under different instructional conditions. Computers in Human Behavior, 96, 211-222 (2019).
10. Moon J. A. A handbook of reflective and experiential learning: Theory and practice. Routledge, New York (2004).
11. Boud, D., Keogh, R., & Walker, D. (Eds.). Reflection: Turning experience into learning. Kogan Page, London (1985).
12. Ullmann, T. D. Automated detection of reflection in texts - A machine learning based approach. Doctoral dissertation, The Open University (2015).
13. Mezirow, J. Transformative dimensions of adult learning. Jossey-Bass, San Francisco, CA (1991).
14. Tsingos, C., Bosnic-Anticevich, S., Lonie, J. M., & Smith, L. A model for assessing reflective practices in pharmacy education. American Journal of Pharmaceutical Education, 79(8) (2015).
15. Ullmann, T. D. Automated analysis of reflection in writing: Validating machine learning approaches. International Journal of Artificial Intelligence in Education, 29(2), pp. 217-257 (2019).
16. Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G., & Dawson, S. Understand students' self-reflections through learning analytics. In: Proceedings of the 8th International Conference on Learning Analytics and Knowledge, pp. 389-398 (2018).
17. Dong, F., Zhang, Y., & Yang, J. Attention-based recurrent convolutional neural network for automatic essay scoring. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pp. 153-162 (2017).
18. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (2018).

19. Pennington, J., Socher, R., & Manning, C. D. Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543 (2014).
20. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018).
21. Cui, Y., Wise, A. F., & Allen, K. L. Developing reflection analytics for health professions education: A multi-dimensional framework to align critical concepts with data features. Computers in Human Behavior, 305-324 (2019).
22. Hmelo-Silver, C. E. Problem-based learning: What and how do students learn?. Educational Psychology Review, 16(3), 235-266 (2004).
23. Van Manen, M. Linking ways of knowing with ways of being practical. Curriculum inquiry, 6(3), 205-228 (1977). doi: 10.1080/03626784.1977.11075533
24. Wong, F. K., Kember, D., Chung, L. Y., & CertEd, L. Y. Assessing the level of student reflection from reflective journals. Journal of Advanced Nursing, 22(1), 48-57 (1995).
25. Poldner, E., Van der Schaaf, M., Simons, P. R. J., Van Tartwijk, J., & Wijngaards, G. Assessing student teachers' reflective writing through quantitative content analysis. European Journal of Teacher Education, 37(3), 348-373 (2014).
26. Zimmerman, B. J., Bonner, S., & Kovach, R. Developing self-regulated learners: Beyond achievement to self-efficacy. American Psychological Association, Washington, D. C. (1996).
27. Cleary, T. J., & Kitsantas, A. Motivation and self-regulated learning influences on middle school mathematics achievement. School Psychology Review, 46(1), 88-107 (2017).
28. Riordan, B., Flor, M., & Pugh, R. How to account for misspellings: Quantifying the benefit of character representations in neural content scoring models. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 116-126 (2019).
29. Saldaña, J. The coding manual for qualitative researchers. Thousand Oaks, CA, Sage (2009).
30. Zhang, W., Yoshida, T., & Tang, X. A comparative study of TF* IDF, LSI and multi-words for text classification. Expert Systems with Applications, 38(3), 2758-2765 (2011).
31. Sultan, M. A., Bethard, S., & Sumner, T. DLS@CU: Sentence similarity from word alignment and semantic vector composition. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 148-153 (2015).
32. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. Scikit-learn: Machine learning in Python. Journal of machine learning research, 12, 2825-2830 (2011).