Improving Automated Evaluation of Formative Assessments with Text Data Augmentation

Keith Cochran[1](✉), Clayton Cohn[2], Nicole Hutchins[2], Gautam Biswas[2], and Peter Hastings[1]

[1] DePaul University, Chicago, IL 60604, USA
kcochr11@depaul.edu
[2] Vanderbilt University, Nashville, TN 37240, USA

Abstract. Formative assessments are an important component of instruction and pedagogy, as they provide students and teachers with insights on how students are progressing in their learning and problem-solving tasks. Most formative assessments are now coded and graded manually, impeding timely interventions that help students overcome difficulties. Automated evaluation of these assessments can facilitate more effective and timely interventions by teachers, allowing them to dynamically discern individual and class trends that they may otherwise miss. State-of-the-art BERT-based models dominate the NLP landscape but require large amounts of training data to attain sufficient classification accuracy and robustness. Unfortunately, educational data sets are often small and unbalanced, limiting any benefits that BERT-like approaches might provide. In this paper, we examine methods for balancing and augmenting training data consisting of students' textual answers from formative assessments, then analyze the impacts in order to improve the accuracy of BERT-based automated evaluations. Our empirical studies show that these techniques consistently outperform models trained on unbalanced and unaugmented data.

Keywords: Data augmentation · Text augmentation · BERT · Formative assessments · Imbalanced data sets · Educational texts · Natural language processing

1 Introduction

The current generation of intelligent learning environments (ILEs) for K-12 students focuses on inquiry, problem-based, game-based, and open-ended learning [10,14,16, for example]. Working on open-ended tasks provides students with

The assessment project described in this article was funded, in part, by the NSF Award # 2017000. The opinions expressed are those of the authors and do not represent views of NSF.

M. M. Rodrigo et al. (Eds.): AIED 2022, LNCS 13355, pp. 390–401, 2022.
https://doi.org/10.1007/978-3-031-11644-5_32

choices in how they develop and pursue their learning and problem-solving processes [22]. Research on ILEs has demonstrated the challenges in framing adaptive support in the context of the specific difficulties that students face as they work on their learning and problem-solving tasks, and develop productive learning strategies [2,21]. Formative assessments have been employed in the learning sciences and education research as interventions that (1) help students learn components of knowledge they need to build and solve larger problem-solving and learning tasks [3] and (2) communicate conceptual understanding of the target domain for self-reflection as well as teacher and environment pedagogical support that aids students' achievement in the context of their current learning [3]. Therefore, formative assessments can help students develop their conceptual understanding of the domain, while supporting their self-assessment and self-regulated learning skills [6,11].

However, formative assessments are often time-consuming to grade [12], limiting the ability to leverage them for *in-time* pedagogical adjustments and feedback. Our long-term goal is to develop robust deep learning-based, natural language processing (NLP) approaches to support rich, in-time formative feedback to students' responses to short answer questions. Formative assessments often go beyond statement-of-fact conceptual knowledge applications, requiring students to reason about causal relations between concepts, explain a scientific process or phenomena, or construct an argument that justifies or negates a particular statement. Simple text-processing methods like keyword matching and templates are often insufficient to uncover the nuanced reasoning in students' short answers to formative assessment questions [13]. Advances in NLP allow us to dive deeper into students' knowledge and reasoning applications, and help students understand the difficulties they face with the instructional material they are being taught. In parallel, they also support teachers in understanding and responding to student difficulties soon after they occur, and before they move on to teach new content. However, issues such as data insufficiency, data imbalance, and lack of variation in student responses limit our ability to apply these advances in a robust and reliable way.

Our approach develops automated text assessments that shed light on students' conceptual knowledge. Educational data sets present several difficulties in NLP because studies typically conducted in classroom environments generate rich data, but the data collection is often limited to about a 100 students at a time. In this paper, we address increasing the effectiveness of NLP evaluation when there is limited and unbalanced training data, as is often the case in educational contexts. We do this by augmenting the training data with generated sentences that share characteristics of the original data. In the rest of this paper, we summarize related work, present our research questions and hypotheses describing the educational context of the formative assessments, and conclude with findings and future work.

2 Background and Research Questions

Transformer-based NLP architectures, such as BERT [8] and GPT-3 [4], are now the industry standard for modeling many NLP tasks. They leverage language knowledge from massive corpora of unlabeled texts via unsupervised pretraining, and they can be fine-tuned on a downstream task with only a fraction of the training instances that would otherwise be required to train a neural network from scratch. However, despite the prevalence of transformer models, many data sets are still too small to effectively fine-tune a model out-of-the-box. There are few areas where this is more apparent than with educational texts in general, and educational assessments in particular.

These texts are also domain-specific, focusing on a wide variety of general areas and specific examples within them. Domain-specific subject matter often includes esoteric jargon that is not well-represented in the canonical corpora that these large transformer models are pretrained on, and there can often be performance degradation when these models are applied to texts whose vocabularies differ considerably from their own [7]. In addition to the issue of educational data sets being non-canonical semantically, they are often non-canonical syntactically as well. Wikipedia, which is used to pre-train both BERT and GPT, is written using proper language syntax. Conversely, many educational texts, such as answers to formative and summative assessments written by children or adolescents, use informal syntax and are written in a much more colloquial manner. This type of text is often incompatible with pre-trained models derived from canonical corpora, as model performance is affected by the quality of data used for training. For example, middle school short answer questions typically use a shallower vocabulary, and this has to be factored into the augmentation techniques used. It is possible to further pre-train BERT with domain-specific corpora, but this also requires large quantities of data. As such, the only practical approach is to select a base model and fine-tune it using labeled data to improve the model's performance.

One salient solution to mitigate the aforementioned issues is *data augmentation*. Data sets once small, imbalanced, and sparsely populated can be made robust by adding instances that are similar in both syntax and semantics. However, hand-crafting these instances can be extremely tedious, so automated approaches are preferable. Data augmentation has been used in areas such as image processing with great effectiveness, e.g., by translating or shifting the images. However in NLP, data augmentation techniques are more complex. A newly generated sentence must retain the same semantic intent as the original sentence. Issues arise when augmented data stray away from the label they are intended to augment. This has led some researchers to assess and label augmented data using experts to ensure correct labeling.

One textual data augmentation technique adds noise in the form of substitution or deletion of words or characters [20]. Another approach uses "back translation" where the data is translated into another language, then translated back, producing alternate ways of saying the same thing [15]. Other forms of data augmentation introduce noise by adding a random character in a word, avoiding the first and last characters of the word. Some methods use random synonym replacements in the form of hypernym (more general) and hyponym (more specific) word replacements using WordNet. Hypernyms have been shown to outperform hyponyms because generalizing a sentence is more likely to preserve the same meaning [9]. BERT uses a masking feature, where a word in the sentence is masked with a special token, and the model tries to predict the masked word. This can serve as another form of augmentation, where a model can be further trained to generate more semantically similar sentences.

In this paper, we examine the benefits of textual data augmentation for evaluating middle school formative assessments (short-answer questions), especially in cases of data scarcity and data imbalance. We compare four different data augmentation techniques: (1) masking using BERT, (2) noise injection, (3) hyponym/hypernym replacement, and (4) oversampling using the existing data. The goal, as discussed, is to provide accurate, timely feedback to students and their teachers. Accordingly, we formulate three research questions:

RQ 1: Does data augmentation improve the classification of student answers? Furthermore, if augmentation is beneficial, is that primarily due to increasing the amount of data, improving the balance between classes, or some combination of both? Our first hypothesis (**H1**) is that both more balanced data and larger amounts of data will improve classification accuracy.

RQ 2: How does the method used for generating new texts affect augmentation performance? Our hypothesis **H2** is that the masking technique will be most effective due to its alignment with BERT. Our expectations for the other three are mixed. In principle, WordNet should provide semantically related substitutes, but its knowledge base is so broad that it may bring in words far outside the learning context.

RQ 3: Do characteristics of the questions and answers affect the effectiveness of data augmentation? For example, some questions may call for fact-based answers. Others may call for descriptions of processes or for causal reasoning that requires the answers to adopt a meaningful structure to produce a correct answer. **H3** proposes that augmenting the data with *wrong answers* will reduce performance because there are such a wide variety of wrong answers for any question. **H4** proposes that augmenting the data with sentences generated from a very small set of examples will also hurt performance due to the limited variability of the samples.

3 The SPICE Curriculum

The formative assessments analyzed in this paper are part of the SPICE (Science Projects Integrating Computation and Engineering) curriculum [23]. This is a three-week, NGSS-aligned unit that challenges students to redesign their schoolyard using appropriate surface materials that meet design constraints and minimize the amount of water runoff after heavy rainfall.

The curriculum (Fig. 1) includes a conceptual modeling unit, where students construct conceptual models of the water runoff phenomenon; then translate it to a computational model of water runoff; and then use the model to solve an engineering design challenge problem, where students construct a playground

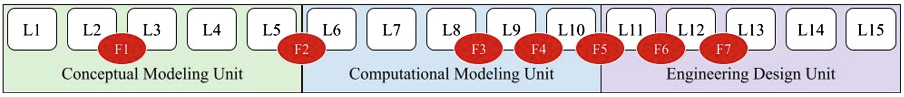


Fig. 1. SPICE curriculum overview.

that adheres to specified constraints [17]. Formative assessments (identified in red) are integrated throughout the curriculum to evaluate students' conceptual understanding in science, computing, and engineering. For this paper, we focus on formative assessment F1 in the conceptual modeling phase.

We leverage evidence-centered design (ECD; [18]) as the overarching framework for assessment development. This process supports our analysis of knowledge construction and problem-solving skill development in the integrated science, computing, and engineering curriculum by linking components of the curriculum and assessments to evidence of students' proficiency with the target knowledge and skills [17]. For instance, students are presented with an incorrect conceptual model and are tasked with (1) identifying and correcting errors and (2) describing positive information presented by the model. These tasks can be linked to key science and engineering practices as described by NGSS [19], including engaging in argument from evidence, and developing and using models, and allow us to evaluate students' science knowledge through its application in model evaluation. In-time analysis of these assessments may allow us to provide key evidence-based, formative feedback to better support students' construction, debugging, and evaluation of their own conceptual models during the curriculum.

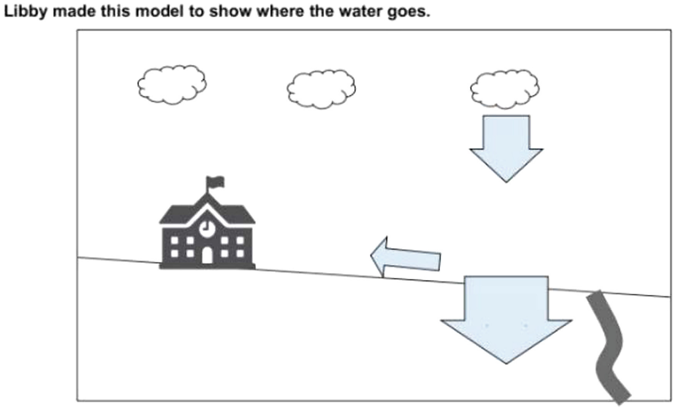
4 Methods

Our exploratory analysis leverages student data collected from a classroom study with 99 6$^{\text{th}}$-grade students in the southeastern United States. The study, conducted in Fall 2019, was led by two experienced science teachers with three university researchers providing additional support in the classroom.

The data set for this study consists of student responses to three separate questions that are based on a fictitious student-constructed visual model shown in Fig. 2. Each question had 95 student responses.[1] The concepts the students must identify for each question are enumerated in Table 1.

1. *What do you think the different sized arrows in Libby's model could mean?* This question has one correct response: the size of the arrows indicates the amount of water. There is only one concept, which evaluates students' understanding of the model representation.
2. *What are two things that you would change about Libby's model to explain where the water goes?*

[1] While there were 99 students in the study, not all students answered each question.



Libby made this model to show where the water goes.

Fig. 2. Libby's model demonstrating where water goes after precipitation.

The focus of this question is on finding errors in the model, explaining the errors, and providing the correct answer. It includes two concepts: the size of the runoff and absorption arrows should sum to the size of the rainfall arrow (conservation of matter), and the direction of the runoff arrow should be pointing downhill. This question evaluates students' knowledge of scientific concepts rather than model representation.

3. *What are two things that Libby's model does a good job of explaining?* Extending the previous question, this question also targets students' ability to observe and evaluate a science model. In this case, although Libby's model contained errors (previous question), the model (1) demonstrates rainfall either is absorbed or becomes runoff, (2) illustrates where water is coming from, and (3) uses arrow size to indicate water amounts. Students are tasked with listing two of these positive model elements and assesses students' knowledge of the scientific concepts as well as their understanding of the model representation.

Table 1. Concepts present in each question.

Question	Concept	Description
1	C1	Arrow size indicates amount of water
2	C2a	Size of runoff and absorption arrows should sum to size of rainfall arrow
2	C2b	Direction of runoff arrow should be pointing downhill
3	C3a	Model demonstrates rainfall either absorbed or becomes runoff
3	C3b	Model illustrates where water is coming from
3	C3c	Model uses arrow size to indicate water amount

Each of the six concepts described above (correct responses for each assessment question) was modeled individually as a binary classification task.

Responses were coded as correct if students identified the concept(s) associated with a specific question, and coded as incorrect otherwise. Note that for questions where there were multiple concepts, the "incorrect" answers include both wrong answers and right answers for other concepts. As previously mentioned, such small educational data sets are often imbalanced. The percentage of the 95 answers for each concept that were labeled as correct is shown in the leftmost data column of Table 2.

4.1 The BERT Model

We used *BERT-base uncased* to classify the student answers because it is widely adopted and is considered state-of-the-art for many NLP tasks. The three sets of student responses for the six different concepts were used for training, validating, and testing the models. For each concept, a separate BERT model was fine-tuned for classification on the training data by adding a single feed-forward layer. We used the micro-F_1 metric as the performance measurement.

In all experiments, the models were trained and evaluated 10 times, with each training iteration using a different seed for the random number generator, which partitions the training and testing instances. During training, the following hyperparameters were used: learning rate 9e-5, batch 12, epochs 2, max sequence 128, train/test split 80/20. Devlin et al. [8] recommend learning rates of 2e-5 to 5e-5, and batch size of 16 or 32, but we chose different values due to data scarcity.

4.2 Baseline Evaluation

For each concept, we evaluated two different baseline models without augmentation or balancing. The *a priori* model simply chose the majority classification for each concept. For our *unaugmented* baseline, we applied BERT in the prototypical way, without data augmentation.

4.3 Augmentation Approach

We chose four textual alteration methods for augmenting the data sets because they are among the leading modern methods at both the word and character level. This gave us a wider sample range to compare and contrast different augmentation methods [1]. Techniques were chosen to minimize the risk of changing semantic intent. WordPiece-level masking is cited as the best augmentation method for classification tasks by Chen et al. [5]. Therefore, our first approach used masking to mask a word in the sentence, then used the BERT model to generate a substitute for the masked word. The second method, noise injection, randomly inserted, deleted or changed a character in the original sentence [9]. The hypo/hypernym method generated sentences by selecting a keyword in the given sentence, and replacing it with both types of related word to generate new candidate sentences [9]. Last, an oversampling method using multiple copies of each instance in the data set was used for augmentation.

The majority label quantity in Table 2 became the **majority quantity of reference** for that particular data set. We first left the data unbalanced from 0x to 1x, then augmented the minority class only by adding $N = (Maj - Min)/5$ sentences[2] at a time until parity was reached. Next, we performed another test by forcing the data to be balanced by removing majority label responses to match the minority level, ensuring parity at each level of augmentation up to 1x. After the data reached 1x, all data sets thereafter were balanced and were augmented in multiples of the majority quantity from 1x to 20x. Initial tests with imbalanced data showed inconsistent results as more augmentation was applied. Additionally, we found empirically that model performance decreased when higher augmentation levels were used over 20x.

5 Results

The high-level view of our results is presented in Table 2. Each row corresponds to a concept. The leftmost data column shows the percentage of the answers for each concept that were originally marked as correct. The next two columns present the baseline results. On the right are the maximum F_1 scores for each concept using one form of data augmentation, and indicating what augmentation quantity level reached that maximum. The highest performance achieved for each concept are shown in bold.

Table 2. Performance (micro-F_1) of baseline vs all augmented models

Concept	%	Baseline		Max Performance	
	Correct	*a priori*	Unaug.	F_1	Aug. Level
C1	89	**0.940**	0.735	0.936	0.6x
C2a	73	0.850	0.757	**0.995**	5x
C2b	33	0.670	0.000	**0.958**	5x
C3a	54	0.700	0.399	**0.873**	8x
C3b	23	0.770	0.000	**0.979**	3x
C3c	40	0.600	0.098	**0.900**	8x

Figure 3 illustrates how balanced and unbalanced data sets affect performance during augmentation. As augmentation increases, the balanced approach shows worse initial performance, however, as augmentation was applied, the balanced approach had a more stable rise in performance. The "balanced" approach (shown by the solid line) forced equality by increasing the minority label as before, but this time, diminishing the majority label such that both had equal representation in the data set.

[2] Here, Maj and Min refer to the number of available sentences from the majority and minority classes.

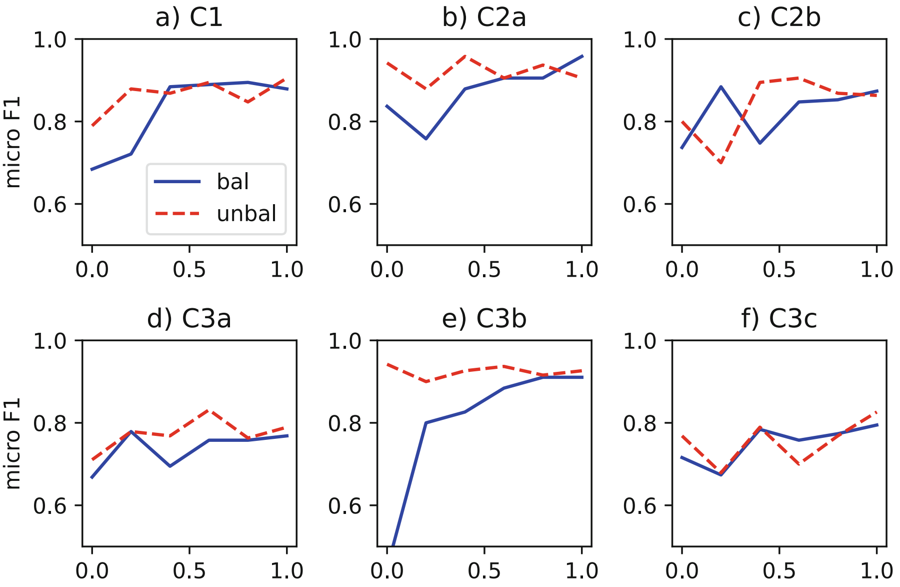


Fig. 3. Model Performance for Balanced vs Unbalanced Data with Augmentation less than 1x the reference majority class.

Figure 4 shows how the performance varied with different augmentation types, masked ("mask"), noise injection ("noise"), hypo-hypernym ("hyp"), and oversampling ("over") and the amount of augmentation for each student response. For each of the concepts, performance improved when augmentation levels from 3x-8x were applied, but tended to fall off slightly with additional augmentation.

6 Discussion

Recall **RQ 1** which asked whether the performance could be improved with an augmented data set. Table 2 shows that data augmentation does improve classification performance over the *a priori* baseline in five of the six concepts, and improves on the unaugmented model baseline in every case. **H1** states that the effect of a balanced data set and larger amounts of data improve classification accuracy. Our results show that balancing the data is vital, and augmentation additionally improves performance. However, there is a limit to how much augmentation can be applied before the model levels off and begins to degrade in performance. Also, for questions with a high percentage of majority label quantities (>90%), guessing the majority label outperforms any model used in our testing. Therefore, **H1** was almost completely confirmed. The only exception was for concept 1, where the majority label represented over 90% of the given responses.

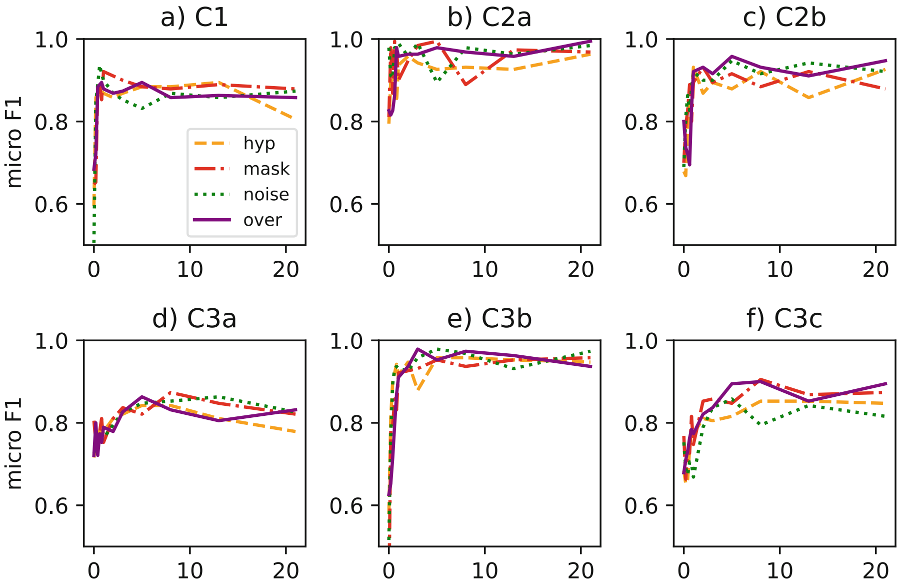


Fig. 4. Model Performance as a Function of Augmentation Level. The x-axis shows the amount of augmentation applied from 0x to 20x.

RQ 2 addresses the four augmentation methods. **H2** predicted that masking would be a clear winner in making the model perform better because of its relation to BERT. Our results in Fig. 4 revealed that performance varies with augmentation method as well as the characteristics of the questions and student responses. No clear winner was evident but a combination of methods may produce better results (currently only an empirical observation). Although **H2** was not supported, stability of performance did vary among the different types of augmentation, so more types of augmentation should be investigated.

RQ 3 speculated that the characteristics of the questions and answers affects model performance. The concept characteristics in our study are that concepts C1, C2a, and C3a are more fact-based answers. The remaining concepts required causal reasoning by the student in order to provide a correct answer. Fact-based concepts do not show a significantly different performance than those requiring causal reasoning. This result means **H3** was not supported.

Figure 3 shows increasing data balance best improves classification accuracy. Once followed up by additional augmentation, significant model improvements can be gained. Some interesting outcomes arose from studying this phenomenon. The model performance improvements in our study varied based on initial data set balance. Data sets that are already close to having an even balance (50%-65%) start out performing in an acceptable range, then increase slightly with applied augmentation to about 8x before falling off. For concepts that originally have close to a two-thirds majority (C2a and C3b), model performance peaked at around 8x.

7 Conclusion

After creating augmented data sets from student responses using four different techniques, then applying them to classify answers using balanced and unbalanced test sets, we found that balancing the data set is the most important feature in achieving improved performance prior to the application of data augmentation. Empirical tests without balance show inconsistent results. However, once balance and augmentation are applied, our experiments showed significant model performance improvements for binary classification of responses. The highest performance improvements occurred when using augmentation levels between 3x to 8x of the quantity of the majority label. Overall, our use of balanced and augmented training sets have generated sufficiently accurate results to support automated grading of formative assessments. However, while promising, our results are still preliminary, and further analysis needs to be conducted.

8 Future Work

Additional augmentation methods (and their combinations) need to be studied to develop models that are more robust for different types of formative assessment questions. After seeing initial ratios of data balance around two-thirds boost performance the most, further investigation is needed. Finally, teachers and education researchers may determine some questions require multiple levels of grading. To grade such answers, we need to train multi-class classifiers or construct hierarchical grading models.

References

1. Bayer, M., Kaufhold, M.A., Reuter, C.: A survey on data augmentation for text classification. arXiv preprint arXiv:2107.03158 (2021)
2. Biswas, G., Segedy, J.R., Bunchongchit, K.: From design to implementation to practice a learning by teaching system: Betty's brain. Int. J. Artif. Intell. Educ. **26**(1), 350–364 (2016)
3. Black, P., Wiliam, D.: Developing the theory of formative assessment. Educ. Assessm. Evaluat. Accountab. **21**, 5–31 (2009)
4. Brown, T.B., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
5. Chen, J., Tam, D., Raffel, C., Bansal, M., Yang, D.: An empirical survey of data augmentation for limited data learning in NLP. arXiv preprint arXiv:2106.07499 (2021)
6. Clark, I.: Formative assessment: assessment is for self-regulated learning. Educ. Psychol. Rev. **24**, 205–249 (2012). https://doi.org/10.1007/s10648-011-9191-6
7. Cohn, C.: BERT Efficacy on Scientific and Medical Datasets: A Systematic Literature Review. DePaul University (2020)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

9. Feng, S.Y., Gangal, V., Kang, D., Mitamura, T., Hovy, E.: GenAug: data augmentation for finetuning text generators. arXiv preprint arXiv:2010.01794 (2020)
10. Geden, M., Emerson, A., Carpenter, D., Rowe, J., Azevedo, R., Lester, J.: Predictive student modeling in game-based learning environments with word embedding representations of reflection. Int. J. Artif. Intell. Educ. **31**(1), 1–23 (2020). https://doi.org/10.1007/s40593-020-00220-4
11. Hattie, J., Timperley, H.: The power of feedback. Rev. Educ. Res. **77**(1), 81–112 (2007). https://doi.org/10.3102/003465430298487
12. Higgins, M., Grant, F., Thompson, P.: Formative assessment: balancing educational effectiveness and resource efficiency. J. Educ. Built Environ. **5**(2), 4–24 (2010). https://doi.org/10.11120/jebe.2010.05020004 https://doi.org/10.11120/jebe.2010.05020004 https://doi.org/10.11120/jebe.2010.05020004
13. Hughes, S.: Automatic Inference of Causal Reasoning Chains from Student Essays. Ph.D. thesis, DePaul University, Chicago (2019). https://via.library.depaul.edu/cdm_etd/19/
14. Käser, T., Schwartz, D.L.: Modeling and analyzing inquiry strategies in open-ended learning environments. Int. J. Artif. Intell. Educ. **30**(3), 504–535 (2020)
15. Liu, P., Wang, X., Xiang, C., Meng, W.: A survey of text data augmentation. In: 2020 International Conference on Computer Communication and Network Security (CCNS), pp. 191–195. IEEE (2020)
16. Luckin, R., du Boulay, B.: Reflections on the Ecolab and the zone of proximal development. Int. J. Artif. Intell. Educ. **26**(1), 416–430 (2015). https://doi.org/10.1007/s40593-015-0072-x
17. McElhaney, K.W., Zhang, N., Basu, S., McBride, E., Biswas, G., Chiu, J.: Using computational modeling to integrate science and engineering curricular activities. In: Gresalfi, M., Horn, I.S. (Eds.). The Interdisciplinarity of the Learning Sciences, 14th International Conference of the Learning Sciences (ICLS) 2020, vol. 3 (2020)
18. Mislevy, R.J., Haertel, G.D.: Implications of evidence-centered design for educational testing. Educational Measurement: Issu. Pract. **25**(4), 6–20 (2006) https://doi.org/10.1111/j.1745-3992.2006.00075.x
19. NGSS: Next Generation Science Standards. For States, By States. The National Academies Press (2013)
20. Wei, J., Zou, K.: EDA: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196 (2019)
21. Winne, Philip H.., Hadwin, Allyson F..: nStudy: tracing and supporting self-regulated learning in the Internet. In: Azevedo, Roger, Aleven, Vincent (eds.) International Handbook of Metacognition and Learning Technologies. SIHE, vol. 28, pp. 293–308. Springer, New York (2013). https://doi.org/10.1007/978-1-4419-5546-3_20
22. Zhang, N., Biswas, G., Hutchins, N.: Measuring and analyzing students' strategic learning behaviors in open-ended learning environments. Int. J. Artif. Intell. Educ. (2021). https://doi.org/10.1007/s40593-021-00275-x
23. Zhang, N., et al.: Studying the interactions between science, engineering, and computational thinking in a learning-by-modeling environment. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12163, pp. 598–609. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52237-7_48