Scaling Mixed-Methods Formative Assessments (mixFA) in Classrooms: A Clustering Pipeline to Identify Student Knowledge

Xinyue Chen and Xu Wang

University of Michigan, Ann Arbor MI 48109, USA
{xinyuech,xwanghci}@umich.edu

Abstract. Formative assessments provide valuable data for teachers to make instructional decisions and help students actively manage their progress and learning. Multiple-choice questions (MCQ) and free-text open-ended questions are typically employed as formative assessments. While MCQs have the benefit of ease of grading and visualizing student answers, they lack capabilities in revealing diverse student ideas and reasoning beyond the options. On the other hand, open-ended tasks and free-text submissions may elicit students' perspectives more comprehensively, though it requires laborious work for instructors to analyze such responses. In this work, we explore the use of mixed-methods formative assessments in a college-level CS class, in which we assign MCQs and ask students to explain their answers. We propose a clustering pipeline to categorize students' free-text explanations leveraging the meta-data the original MCQs provide. We find that using students' choices in MCQs to resolve co-reference in their explanations and adding students' choices as features significantly improve clustering performance. Moreover, our work demonstrates that providing structures in the data collection process improves the clustering of free-text responses without making changes to the algorithm.

Keywords: Formative Assessments · Self-Explanations · Clustering Pipeline

1 Introduction

College classes witness high enrollment in recent years [13]. Especially in computer science, students in introductory courses are from increasingly diverse backgrounds [1]. This introduces difficulties for instructors to accurately and efficiently predict students' knowledge and monitor student progress to plan for and adjust their instruction [14, 22]. In-class formative assessments, in the format of multiple-choice questions (MCQ) or open-ended questions (OEQ), are often employed by instructors to identify students' strengths and weaknesses. As an example, during a lecture, an instructor may use MCQs to probe into students' understanding of concepts and visualize student options in real-time [36]. In other cases, instructors may use OEQs and walk around the classroom to sample students' answers and prompt the class to discuss further [7].

Although MCQs have the benefit of ease of grading and help instructors quickly visualize student answers, prior work has raised concerns on whether the

options experts designed could correctly and comprehensively reflect students' understanding and misconceptions [19, 21]. Some studies have shown that learners may be over-tested by MCQ because they can select the right answer even when they are not able to complete the task [10]. Moreover, instructors could gain little insights into the reasoning behind students' choices [21]. On the other hand, OEQs have the benefit of revealing students' ideas and reasoning behind a problem [21]. However, using OEQs as formative assessments lacks the immediacy for instructors to identify students' weaknesses and monitor their progress, as analyzing a large amount of textual data is laborious [28,32]. To solve this problem, researchers have explored Natural Language Processing (NLP) methods to detect the common misconceptions in students' textual responses [26, 29, 33], however, it remains a challenging task for several reasons. First, it is difficult to parse the contextual information in students' answers, e.g. domain-specific terms and abbreviations, and incomplete sentences; Second, students' answers often have nuanced differences in the meaning they convey, but existing studies focus on detecting right answers from wrong rather than capture diverse students' perspectives. Third, although we have seen well-performed domain-specific classification models in short-answer grading, the generalizability across question topics and disciplines is unsatisfactory [29].

In this work, we explore the use of *mixed-methods formative assessments (mixFA)* to identify students' knowledge. Specifically, in a college-level user interface development class with 373 students, we assigned MCQs and ask students to explain their answers. We created a mixFA dataset with labels of students' ideas as shown in their explanations. We propose a clustering pipeline to categorize students' free-text explanations leveraging the meta-data the original MCQs provide. We see several benefits of using mixFA. First, mixFA elicits in-depth student reasoning and diverse student ideas compared to using MCQs alone. Second, the clustering pipeline can quickly and effectively cluster students' free-text explanations. We find using students' choices in MCQs to resolve co-reference in their explanations and adding students' choices as features significantly improve clustering performance.

We present a case study where providing structures in the data collection process improves the clustering of free-text student responses without making changes to the algorithm. We discuss the implications on collecting meta-data and improving feature representation as our community makes improvements on short answer clustering and classification problems. Through a qualitative error analysis of the clustering outcome, we surface the need to give instructors more control over the clustering setup, e.g., providing input for the algorithm to improve and being able to explore and rectify clustering results. We discuss implications on building human-in-the-loop interfaces to invite instructor input and allow for more versatile NLP-powered short answer clustering and classification pipelines. We suggest that mixFA could support instructors in identifying students' knowledge and monitoring student progress in a way that achieves quality and scale at the same time.

2 Related Work

In this section, we discuss relevant literature on how formative assessments can be used in classrooms to help instructors with decision making and prior machine learning-powered methods to identify students' knowledge and ideas.

2.1 Formative Assessments in Supporting Teaching and Learning

Decades of research has shown the benefit of using formative assessments to facilitate student learning and help instructors identify students' strengths and weaknesses to adjust their teaching [4, 24]. In the process, it is critical for instructors to analyze student responses and translate the insights to help with their instructional decision-making [16]. Commonly used formative assessments could take the form of multiple-choice questions (MCQ) or open-ended questions (OEQ). While MCQs have the benefit of ease of grading, they may not elicit students' prior knowledge and ideas comprehensively [12]. On the other hand, while OEQs are better positioned to capture diverse student ideas, they do not provide immediacy for instructors to visualize student answers [4]. Research has shown that integrating self-explanations into MCQs could improve students' learning of complex concepts and skills, and help students develop meta-cognitive skills [5]. In our study, we explore the use of *mixFA*, combining MCQs with self-explanation prompts. We investigate whether mixFA could elicit diverse student reasoning and ideas beyond the options in MCQ and offer opportunities for automatically clustering student ideas and reasoning. This can support downstream educational applications, including supporting automatic short answer grading [27], crowdsourcing explanations for future students [34], and generating high quality questions leveraging natural student mistakes [32, 33].

2.2 Automatic Methods for Identifying Students' Prior Knowledge and Misconceptions

Prior work has explored machine learning techniques to detect students' prior knowledge in short-answer textual responses [8, 17, 26]. Michalenko et al. developed a probabilistic model to differentiate students' correct and wrong answers [17]. Other work proposed NLP models to cluster students' short answers, with a focus on programming tasks that provide more structured features than free text [25, 26]. More recent work applied pre-trained language models, such as BERT on short answer grading [6, 18, 30]. They found that transformers improved the accuracy of automatic grading results. We summarize the following limitations in prior work: 1) Most existing techniques in the space of classifying and clustering student free-text responses have a focus on detecting correct and incorrect answers. However, in a formative assessment setting, instructors have the desire to identify diverse student ideas and reasoning to plan for and adjust their teaching [4,24]. 2) Although we have seen successes with recent short answer grading techniques, the performance remains inconsistent across data sets [9]. Domain-specific models require substantial efforts on data annotation, whereas domain-general models also require abundant data input [35]. Nuanced meanings conveyed in students' short answers are hard to be captured with existing

approaches [26]. In this work, through collecting the mixFA response dataset, we investigate whether structured student responses allow for the development of novel clustering pipelines to help instructors identify students' knowledge and misconceptions using mixFA.

3 MixFA Response Dataset

3.1 Data Collection

We explored the use of *mixFA* and collected a response dataset in a college-level introductory Human-Computer Interaction course with 373 students at the University of Michigan. Through discussion with two of the course instructors, we designed 7 Multiple-choice questions (MCQ) on topics including ideation, prototyping, think-aloud protocols, and universal and accessible design principles. Each MCQ offers 4-5 different options for students to choose from. The options were designed based on both instructors' predictions of students' prior knowledge and past students' mistakes. The MCQs were used in mixFA, in which students were also asked to explain their answers. The class was offered in the fall of 2021. The study was IRB approved, and 373 students in the class consented to have their data collected. At the end of the class, we collected 987 mixFA responses (with student MCQ choices and explanations).

3.2 Data Preparation

Since our goal is to investigate whether mixFA can elicit student reasoning and misconceptions behind their choices, we developed a coding scheme for each question to annotate unique student ideas or misconceptions emerging from their mixFA explanations. For each of the 7 questions, one author did the initial coding of the mixFA explanations. In the first step, answers that did not contain explanatory information (e.g., "refer to the slides"; "In lecture"; "Yes/No") were coded as "Non-informative". There were 284 student responses labeled as "non-informative" and excluded for further analysis. Two authors then did axial coding based on the initial codes. In this process, we made sure all initial codes with similar meanings were merged and determined whether a code is a correct understanding or a misconception. We then developed a codebook for each question which showed unique student ideas emerging from the mixFA explanations. Two authors used the codebook to code 10% of data for each question in the mixFA dataset and achieved an average Cohen's kappa [15] of 0.91. One author then coded the whole mixFA dataset. Table 1 displays examples with initial and final codes after merging.

Student Explanations	**Initial Code**	**Final Code**
You want to get a lot of ideas first and then judge them	Get idea first and then judge them later	Get ideas first (without judgment) and then evaluate/narrow them down later.
Evaluation of quality is certainly more required for the next step in the process	Evaluation of the quality in the next step	Get ideas first (without judgment) and then evaluate/narrow them down later.

Table 1: An example of merging similar initial codes to a final code.

Following the data annotation process, we built the mixFA response dataset with each student explanation labeled with a code representing a unique idea. The labels are used as the ground truth for our subsequent clustering and classification experiments. One thing to highlight here is that our ultimate goal is to support instructors using mixFA to identify diverse students' knowledge and misconceptions. So we tried our best to retain the meaning in students' explanations and made nuanced distinctions between codes in our coding process. Some codes may share common keywords but they demonstrate different specificity and levels of understanding from the students. For example, *"Block-based programming is easier because dragging is easier than typing for people with motor disabilities"* and *"Block-based programming is easier or more accessible"* are treated as different codes, since the former one displays extra reasoning, and both codes are misconceptions. The dataset includes 703 annotated free-text self-explanations in response to 7 multiple-choice questions. The dataset and the coding manual can be downloaded at this link [1].

4 Methods: A Clustering Pipeline for Identifying Students' Knowledge

To help instructors identify diverse student ideas and reasoning from students' self-explanations in mixFA, we develop a clustering pipeline. The novelty of the clustering pipeline lies in applying meta-level data that mixFA responses provide. Specifically, we use the original MCQ options to resolve the co-reference in students' explanations and use the MCQ answer as an additional feature.

4.1 Co-reference Resolution

One challenge presented in short answer grading is that incomplete sentences are common [9]. Similarly, in our mixFA dataset, student explanations often rely on contextual information in the question itself. For example, students may use pronouns or abbreviations to refer to the entities in the original MCQ options. Thus we applied co-reference resolution to contextualize students' explanations. Specifically, we used the NeuralCoref pipeline in SpaCy to resolve co-references with the combined option and explanation as input [11]. We then split the output to extract the resolved explanations. Here we present an example, before co-reference resolution: "It is an iterative process.", and after co-reference resolution: "The transition between lo-fi and high-fi prototyping is an iterative process." More examples are shown in Table 4

4.2 Data Representation

We used sentence-BERT [23] to represent the textual data. Sentence-BERT is a state-of-the-art method for sentence embeddings. It utilized siamese and triplet network structures to derive semantically meaningful sentence embeddings. Prior work showed that sentence-BERT performed well on short answer grading tasks in an educational context [6,20], with better performance on clustering tasks than

[1] https://github.com/UM-Lifelong-Learning-Lab/AIED2022-MixFA-dataset.

alternative GloVe and BERT embeddings [23]. We also tried Word2Vec, GloVe, BERT, sentence-BERT to represent student text answers and found sentence-BERT to be the best by comparing the clustering resutls with manual labels.

We extracted students' answers in the corresponding MCQ as an additional feature since students' MCQ answers represent their prior knowledge [21]. For example, for Question 3 as shown in Table 3, students explain the transition between low-fidelity and high-fidelity prototypes when they select option A while focusing on the benefit of the low-fidelity prototype when they select option B. In our dataset, students can have up to 14 different combinations of option selection since some MCQs used were select all that apply questions. To construct the feature space, we combined the feature column of students' MCQ answers with the vectorized explanation using sentence-BERT.

4.3 Clustering

We used the agglomerative clustering method with euclidean distance measure and average linkage provided in Scikit-Learn to cluster students' explanations leveraging the feature representation presented above. Agglomerative clustering is a bottom-up algorithm that treats each data point as a singleton cluster at the outset and then successively agglomerates pairs of clusters until all clusters have been merged into a single cluster that contains all data. We adopted this approach since it resembles instructors' natural process of discovering and merging different student ideas. In this study, we examine whether providing structure in the data collection process could improve the clustering of free-text student responses. We set the number of clusters to be the same as the number of codes extracted from the annotation process. We then evaluate the clustering outcome by comparing the results with our manual labels using Adjusted Mutual Information Score [31] Adjusted mutual information score (AMI) is a commonly used metric for comparing clustering outcomes and it corrects the effect of the agreement solely due to chance between clustering algorithms [31]. We also use the Silhouette Coefficient score to evaluate the density of the clusters [3].

5 Findings

In this section, we report the performance of the clustering pipeline with baseline models. We also report findings on an in-depth error analysis of the clustering outcome to suggest future pathways for more effective clustering of students' free-text answers.

5.1 Experiment Results

We use the mixFA dataset, as shown in the public link [1]. We run the clustering algorithm separately for each of the 7 questions. There are four experimental setups with different feature representations: 1) Sentence-BERT only; 2) Sentence-BERT applied after co-reference resolution (Resolved-SBERT); 3) Sentence-BERT plus MCQ options as a column feature (Option-SBERT); 4) Sentence-BERT applied after co-reference resolution plus MCQ options as a column feature (Resolved-Option-SBERT).

Meta-Level Data from MCQ Improves the Clustering Outcome Table 2 shows the AMI scores for the four experimental setups for each of the 7 questions. We applied Anova one-way analysis with Dunn's posthoc pairwise test. The column "p" shows the p-value of each model compared with the baseline model in Dunn's test. We see marginally significant improvement in Adjusted Mutual Information score with Resolved-SBERT(Ave. AMI = 0.30, p <0.1) and significant improvement with Option-SBERT conditions(Ave. AMI = 0.34, p <0.05) . Maximum improvement is obtained when using both resolved explanations and the MCQ options as a column feature (Ave. AMI = 0.42, p <0.01). This suggests that our proposed feature space with co-reference resolution and MCQ options improves the data representation.

Model	Questions	Q1	Q2	Q3	Q4	Q5	Q6	Q7	***Ave.***	***p***
	Clusters	18	22	14	16	17	18	13		
Baseline	AMI	0.14	0.30	0.19	0.11	0.3	0.16	0.24	0.21	
Resolved-SBERT	AMI	0.23	0.40	0.32	0.19	0.37	0.27	0.39	0.30	0.08*
Option-SBERT	AMI	0.18	0.41	0.36	0.21	0.33	0.30	0.51	0.34	0.04**
Resolved-Option-SBERT	AMI	0.29	0.44	0.38	0.29	0.41	0.36	**0.61**	**0.41**	0.003***

Table 2: Adjusted Mutual Information score(AMI) for the four experimental setups with different feature representations. AMI score improved significantly in Resolved-SBERT, Option-SBERT, and Resolved-Option-SBERT compared with the Baseline.

Tradeoff Between Capturing Nuanced Differences in Student Answers and Achieving Better Clustering Outcomes In the experiments, we set the number of clusters to be the same as the number of manual labels shown in our dataset [1], which gives us a relatively large number of clusters (average clusters = 17) under each question. Therefore, the lower distance-thresholds for hierarchical clustering will increase the possibility that student explanations with a certain degree of similarity are not merged, causing the errors. We present evidence that reducing the number of clusters may increase the AMI score and the Silhouette score. However, that will leave some unique student ideas, and nuances between student explanations uncaptured. Figure 1 shows the average AMI score across 7 questions when changing the number of clusters. We can see that for the Resolved-Option-BERT setup, the AMI score is peaked when N (number of clusters) = 13. Figure 2 shows the average Silhouette score across 7 questions when changing the number of clusters. A general trend is that the silhouette score is higher when there are fewer clusters. This is understandable because student answers may appear linguistically similar but convey different meanings.

This set of experiments demonstrates that if our goal is to capture diverse student ideas in a formative assessment scenario for instructors to understand student's knowledge and reasoning, especially the subtle differences in student answers, optimizing for existing ML metrics (such as AMI or Silhouette score) may not be sufficient.

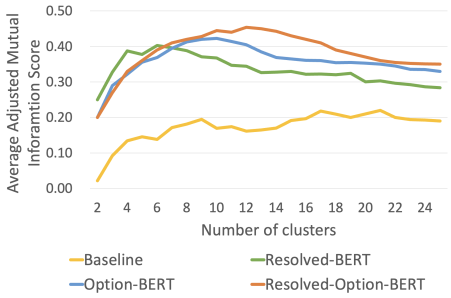

Fig. 1: Average AMI changes with the number of clusters. For the Resolved-Option-BERT setup, AMI peaks when N= 13, <the number of manual labels.

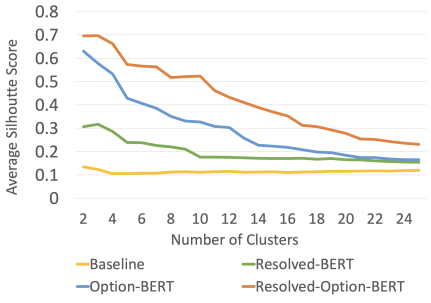

Fig. 2: The average Silhouette score increases as the number of clusters decreases. Resolved-Option-BERT setup has the overall best performance.

5.2 Qualitative Assessment

We performed an in-depth qualitative assessment of the clustering outcome to see how the new feature representation influenced the result and where the errors came from. We summarize the drawbacks of the clustering pipeline and propose future improvement ideas. We use Question 3 as an example, as shown in Table 3.

Which of the following is NOT correct about the relationship between low-fidelity versus high-fidelity prototypes? Select all that apply. (Correct Answers: B,C)
A. It is always better to first do a low-fidelity prototype versus a high fidelity prototype because we need to know the basics of user interaction
B. Lo-fi prototypes, if done well, could give us everything we need to understand user interactions with the system.
C. The transition between lo-fi and high-fi prototyping is a linear process.
D. Lo-fi prototypes could provide us with valuable data and help us evaluate high-level characteristics of the system that could inform us on how to build a high-fi prototype.

Table 3: An example question in the mixFA dataset (Question 3). Students are provided with a text field following this question to explain their answers.

Benefits of the New Data Representation with Co-reference Resolution and MCQ Option Column Feature The co-reference resolution step successfully helps complete students' sentences. Table 4 shows examples where entities in students' explanations are successfully replaced and enriched with co-reference resolution. We found that adding the MCQ options as a column feature had mixed effects on the clustering outcome. On the one hand, the option feature helps when student explanations are aligned with the original options, eg., the second example shown in Table 4. However, when student explanations are widely disparate, the option feature is distracting, eg., the first example shown in Table 4 .

Error Analysis One source of error we observed was that two labels were clustered together. This was often due to the fact that student answers present similar linguistic features, however, when we analyze them qualitatively, they

	Options that students chose	Explanations	Resolved Explanations
1	The transition between lo-fi and high-fi prototyping is a linear process.	It is an iterative process.	The transition between lo-fi and high-fi prototyping is an iterative process.
2	Lo-fi prototypes, if done well, could give us everything we need to understand user interactions with the system.	It wouldn't give us everything we need to know	Lo-fi prototypes wouldn't give us everything us need to know.

Table 4: Example of Successful Co-reference Resolution. The subject in the student's explanation was correctly replaced with domain-specific keywords.

demonstrate subtle differences in student understanding. For example, students' explanations "Low-fi are important for an initial part of the prototyping process" and "Do Lo-fi at first helps gather data to build hi-fi" were grouped in one cluster as they were similar to some extent. However, in our manual coding, we take these as two different students' perspectives, "It is helpful to first do lo-fi first." and "Do Lo-fi first could help with hi-fi.", because the latter one is more specific about the relationship between lo-fi and hi-fi. On the contrary, another type of error is that students' explanations in one label were distributed into two clusters due to the length or quality of the explanations. For example, for the label "Lo-fi can not represent everything", students' simple answers such as *"Not everything"* were placed into one cluster, while other explanations with higher specificity such as *"Lo-fi prototypes intentionally exclude some of the details about how the app works"* were placed into a different cluster.

Another main source of error was that incorrect and correct explanations with similar linguistic features were wrongly clustered together, e.g., students' explanation *"It is always better to do low-fi first"* and *"It's not necessarily true that it's always better to do low-fi first"* were wrongly clustered together. Since it is critical to recognize the polarity and sentiment in student answers, future work could incorporate additional features to highlight such tendencies during data representation. Besides, the linguistic distance between MCQ options, and the level of student knowledge they represent could serve as additional features. For example, some options are partially incorrect, whereas others are completely wrong. We also observe cases where co-reference resolution doesn't work well. This may happen when the option sentence has complex structures with multiple entities.

These errors point to design ideas for giving instructors more control in the process and interaction with the clustering or classification algorithm. First, in the mixFA dataset, the explanations for different questions possess varying properties, e.g., to what extent student explanations target the options in the MCQ. We can give instructors more control to decide what data representations to use, adjust the number of clusters, and determine the threshold for clustering depending on the nuanced level they want to get at. Second, instructor input on keywords, synonyms, and opposing arguments could help correct many of the errors we have seen in the experiments. Lastly, when clustering outcomes are not ideal, instructors need to have the freedom to freely explore and rectify the clustering result.

5.3 Validation of the New Feature Representation

We applied a supervised learning approach to examine how the new data representation supports classification compared to existing approaches on short answer grading. Specifically, with the feature representation of the Resolved-Option-BERT setup, we trained logistic regression classifiers and evaluated the classifiers through 10-fold cross-validation. The results are shown in Table 5. In comparison to a recent study [6] which uses SBERT for student answer classification (SBERT accuracy, 0.621), our setup reaches a higher level of accuracy (Resolved-option-SBERT, 0.661). This offers triangulation that the meta-level data provided by mixFA improves the data representation in students' free-text explanations.

	SBERT-baseline	**Resolved-SBERT**	**Option-SBERT**	**Resolved-Option-SBERT**	**Condor, 2021 [6]**
Accuracy	0.587	0.612	0.629	**0.661**	0.621
AMI	0.245	0.298	0.358	0.422	—

Table 5: Accuracy of the classifiers built on the mixFA dataset with a 10-fold cross validation. This offers triangulation that the meta-level data provided by mixFA improves the data representation in students' free-text explanations.

6 Discussion and Conclusion

In this work, we contribute the mixFA dataset which contains students' answers to MCQ questions and their free-text explanations. We then propose a clustering pipeline that improves the vectorization of students' free-text explanations using the meta-level data the corresponding MCQs provide. Our findings show that MCQ options could be used to resolve co-references in students' free-text answers and their MCQ choices provide additional context for clustering. We demonstrate that the clustering pipeline with co-reference resolution and the choice information significantly outperforms the baseline setup with sentence-BERT only. We show a case study where providing structures in the data collection process improves the clustering of free-text student responses without making changes to the algorithm. Besides, our findings show the trade-offs between capturing nuanced differences in students answers and optimizing for metrics such as the AMI and the Sillhoutte scores. Future studies in the space need to devise and use metrics that are aligned with instructional goals.

We present a qualitative error analysis which points to failure cases of the proposed clustering pipeline. We discuss the design implications on building a human-in-the-loop interface where instructors control the clustering setup and provide input to improve the outcomes [2]. For example, instructors may experiment with alternative data representations, choose when and how to use meta-level data, provide keywords and synonyms, specify opposing arguments, and rectify clustering mistakes.

In conclusion, our study shows that mixFA is a viable approach for eliciting diverse and nuanced student ideas and reasoning, while at the same time instructors can use the clustering pipeline to quickly examine students' knowledge.

References

1. Alhazmi, S., Hamilton, M., Thevathayan, C.: Cs for all: Catering to diversity of master's students through assignment choices. In: Proceedings of the 49th ACM Technical Symposium on Computer Science Education. pp. 38–43 (2018)
2. Amershi, S., Cakmak, M., Knox, W.B., Kulesza, T.: Power to the people: The role of humans in interactive machine learning. Ai Magazine **35**(4), 105–120 (2014)
3. Aranganayagi, S., Thangavel, K.: Clustering categorical data using silhouette coefficient as a relocating measure. In: International conference on computational intelligence and multimedia applications (ICCIMA 2007). vol. 2, pp. 13–17 (2007)
4. Bennett, R.E.: Formative assessment: A critical review. Assessment in education: principles, policy & practice **18**(1), 5–25 (2011)
5. Chung, C.Y., Hsiao, I.H.: Examining the effect of self-explanations in distributed self-assessment. In: European Conference on Technology Enhanced Learning. pp. 149–162. Springer (2021)
6. Condor, A., Litster, M., Pardos, Z.: Automatic short answer grading with sbert on out-of-sample questions. International Educational Data Mining Society (2021)
7. Crouch, C.H., Mazur, E.: Peer instruction: Ten years of experience and results. American journal of physics **69**(9), 970–977 (2001)
8. Feldman, M.Q., Cho, J.Y., Ong, M., Gulwani, S., Popović, Z., Andersen, E.: Automatic diagnosis of students' misconceptions in k-8 mathematics. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (2018)
9. Galhardi, L.B., Brancher, J.D.: Machine learning approach for automatic short answer grading: A systematic review. In: Ibero-american conference on artificial intelligence. pp. 380–391. Springer (2018)
10. Harrison, C.J., Könings, K.D., Schuwirth, L.W., Wass, V., Van der Vleuten, C.P.: Changing the culture of assessment: the dominance of the summative assessment paradigm. BMC medical education **17**(1), 1–14 (2017)
11. Huggingface: Huggingface/neuralcoref: fast coreference resolution in spacy with neural networks, `https://github.com/huggingface/neuralcoref`
12. Kanli, U.: Using a two-tier test to analyse students' and teachers' alternative concepts in astronomy. Science Education International **26**(2), 148–165 (2015)
13. Kara, E., Tonin, M., Vlassopoulos, M.: Class size effects in higher education: Differences across stem and non-stem fields. Economics of education review **82** (2021)
14. Karataş, P., Karaman, A.C.: Challenges faced by novice language teachers: Support, identity, and pedagogy in the initial years of teaching. The International Journal of Research in Teacher Education **4**(3), 10–23 (2013)
15. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics pp. 159–174 (1977)
16. Mandinach, E.B., Gummer, E.S., Muller, R.D.: The complexities of integrating data-driven decision making into professional preparation in schools of education: It's harder than you think. In: Alexandria, VA: CNA Analysis & Solutions (2011)
17. Michalenko, J.J., Lan, A.S., Baraniuk, R.G.: Data-Mining Textual Responses to Uncover Misconception Patterns. In: Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale. L@S '17, New York, NY, USA (2017)
18. Nandini, V., Maheswari, P.U.: Automatic assessment of descriptive answers in online examination system using semantic relational features. The Journal of Supercomputing **76**(6), 4430–4448 (2020)
19. Nathan, M.J., Petrosino, A.: Expert blind spot among preservice teachers. American educational research journal **40**(4), 905–928 (2003)

20. Ndukwe, I.G., Amadi, C.E., Nkomo, L.M., Daniel, B.K.: Automatic grading system using sentence-bert network. In: International Conference on Artificial Intelligence in Education. pp. 224–227. Springer (2020)
21. Polat, M.: Analysis of multiple-choice versus open-ended questions in language tests according to different cognitive domain levels. Novitas-ROYAL (Research on Youth and Language) **14**(2), 76–96 (2020)
22. Qian, Y., Lehman, J.: Students' misconceptions and other difficulties in introductory programming: A literature review. ACM Transactions on Computing Education (TOCE) **18**(1), 1–24 (2017)
23. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
24. Schildkamp, K., van der Kleij, F.M., Heitink, M.C., Kippers, W.B., Veldkamp, B.P.: Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. International Journal of Educational Research **103** (2020)
25. Shi, Y., Mao, T., Barnes, T., Chi, M., Price, T.W.: More with less: Exploring how to use deep learning effectively through semi-supervised learning for automatic bug detection in student code. In: In Proceedings of the 14th International Conference on Educational Data Mining (EDM) 2021 (2021)
26. Shi, Y., Shah, K., Wang, W., Marwan, S., Penmetsa, P., Price, T.: Toward Semi-Automatic Misconception Discovery Using Code Embeddings. In: LAK21: 11th International Learning Analytics and Knowledge Conference. pp. 606–612 (2021)
27. Singh, A., Karayev, S., Gutowski, K., Abbeel, P.: Gradescope: a fast, flexible, and fair system for scalable assessment of handwritten work. In: Proceedings of the fourth (2017) acm conference on learning@ scale. pp. 81–88 (2017)
28. Sirkiä, T., Sorva, J.: Exploring programming misconceptions: an analysis of student mistakes in visual program simulation exercises. In: Proceedings of the 12th International Conference on Computing Education Research. pp. 19–28 (2012)
29. Sung, C., Dhamecha, T.I., Mukhi, N.: Improving short answer grading using transformer-based pre-training. In: International Conference on Artificial Intelligence in Education. pp. 469–481. Springer (2019)
30. Uto, M., Uchida, Y.: Automated short-answer grading using deep neural networks and item response theory. In: International Conference on Artificial Intelligence in Education. pp. 334–339. Springer (2020)
31. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. The Journal of Machine Learning Research **11**, 2837–2854 (2010)
32. Wang, X., Rose, C., Koedinger, K.: Seeing beyond expert blind spots: Online learning design for scale and quality. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–14 (2021)
33. Wang, X., Talluri, S.T., Rose, C., Koedinger, K.: Upgrade: Sourcing student open-ended solutions to create scalable learning opportunities. In: Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale. pp. 1–10 (2019)
34. Williams, J.J., Kim, J., Rafferty, A., Maldonado, S., Gajos, K.Z., Lasecki, W.S., Heffernan, N.: Axis: Generating explanations at scale with learnersourcing and machine learning. In: Proceedings of the Third (2016) ACM Conference on Learning@ Scale. pp. 379–388 (2016)
35. Zhang, L., Huang, Y., Yang, X., Yu, S., Zhuang, F.: An automatic short-answer grading model for semi-open-ended questions. Interactive learning environments **30**(1), 177–190 (2022)
36. Zou, D., Xie, H.: Flipping an english writing class with technology-enhanced just-in-time teaching and peer instruction. Interactive Learning Environments (2019)