Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Predicting and Analyzing Students' Higher-Order Questions in Collaborative Problem-Solving

Shan ZHANG[a*], Toni V. EARLE-RANDELL[a], Qian SHEN[a], Anthony F. BOTELHO[a], Maya ISRAEL[a], Kristy Elizabeth BOYER[a], Collin F. LYNCH[b] & Eric WIEBE[b]
[a]*University of Florida, USA*
[b]*North Carolina State University, USA*
*zhangshan@ufl.edu

Abstract: Question-asking is a crucial learning and teaching approach. It reveals different levels of students' understanding, application, and potential misconceptions. Previous studies have categorized question types into higher and lower orders, finding positive and significant associations between higher-order questions and students' critical thinking ability and their learning outcomes in different learning contexts. However, the diversity of higher-order questions, especially in collaborative learning environments, has left open the question of how they may be different from other types of dialogue that emerge from students' conversations. To address these questions, our study utilized natural language processing techniques to build a model and investigate the characteristics of students' higher-order questions. We interpreted these questions using Bloom's taxonomy, and our results reveal three types of higher-order questions during collaborative problem-solving. Students often use "Why", "How" and "What If" questions to 1) understand the reason and thought process behind their partners' actions; 2) explore and analyze the project by pinpointing the problem; and 3) propose and evaluate ideas or alternative solutions. In addition, we found dialogue labeled 'Social', 'Question - other', 'Directed at Agent', and 'Confusion/Help Seeking' shows similar underlying patterns to higher-order questions. Our findings provide insight into the different scenarios driving students' higher-order questions and inform the design of adaptive systems to deliver personalized feedback based on students' questions.

Keywords: Higher-Order Questions; Dialogue Modeling; Collaborative Learning

1. Introduction

When students face a challenge or find gaps in their understanding, they seek help from their peers or teachers. Oftentimes, this help-seeking presents in the form of a question, looking for the knowledge they are missing to solve a problem (Boldero & Fallon, 1995; Puustinen, 1998). The types of questions that learners ask can indicate the type of knowledge they are looking for. Thus determining what kind of questions they are asking can reveal the kind of understanding that students already have and their goals when looking for help (Qayyum, 2018). This can, in turn, provide us with a deeper understanding of how engaged students are in the activity (Karabenick & Knapp, 1991).

As students pose questions, they not only clarify their current understanding but also explore procedural skills and conceptual knowledge, cultivating their higher-order thinking skills. Procedural skill is tied to specific problems and the ability to execute the steps to solve them, while conceptual understanding is more implicit, focusing on the understanding of the principles within a domain (Hiebert & Lefevre, 1986; Rittle-Johnson, et al., 2001). Bloom's Taxonomy (Anderson & Krathwohl, 2001) provides a framework for us to describe and characterize levels of student understanding, and these levels can be categorized into low-

order and higher-order thinking (Hopper, 2006). Lower-order thinking (represented as Knowledge, Comprehension, and Application in Bloom's Taxonomy) tends to present as closed questions with concrete answers, and while recall and comprehension are necessary skills on a learner's toolkit, higher-order thinking encourages critical analysis and evaluation of concepts, asking 'Why?' (Achmad & Utami, 2023; Khan & Inamullah, 2011). Higher-order questions tend to be open-ended questions that require implicit conceptual knowledge and engage students in the higher-order thinking skills of Analysis, Synthesis, and Evaluation (Hopper, 2006). The act of forming these questions inherently involves higher-order thinking as students get engaged with the materials at a deep level, exploring what they know, and understanding and evaluating how and why the knowledge is constructed and applied

Collaborative learning provides a unique opportunity to examine students' learning processes as they interact with both their peers and learning materials. In such a paradigm, students can communicate with their peers through question-asking and other forms of dialogue to seek knowledge when filling a knowledge gap. While lecture-centered approaches to learning typically do not foster critical thinking skills (Bustami et al., 2018), student-centered approaches like collaborative learning environments allow students to engage in discussions and group activities that challenge them and provide them the freedom to build those higher-order thinking skills (Alharbi et al., 2022). This research builds upon prior work that investigates how young students collaborate to solve a coding task, focusing on exploratory talk and higher-order dialogue (Earle-Randell et al., 2023). In the previous study, learners worked together on one computer using "Pair programming" (Williams, et al., 2002), in which the students take turns using the controls, with the driver controlling the mouse and keyboard, and the navigator contributing their ideas and helping the driver complete the task.

The primary objective of our study is to investigate the underlying language patterns of students' higher-order questions during this collaborative coding task. We analyzed students' dialogue using machine learning techniques to investigate the higher-order questions asked between learners through the following research questions:

- RQ1: How well can we predict higher-order questions in problem-solving collaborative discourse during a pair programming task?
- RQ2: What clusters of higher-order questions emerge from students' discourse in this task?

Analyzing language patterns during collaborative problem-solving and categorizing students' higher-order questions would allow us to develop a deeper understanding of the processes behind higher-order thinking. A robust higher-order question model could not only save teachers' time manually identifying students' higher-order questions but also assist in the development of adaptive learning systems that deliver individualized learning experiences to students by adjusting automated feedback to suit their questions and needs.

2. Literature Review

For its ability to describe and characterize different levels of students' conceptualization, Bloom's Taxonomy (Anderson & Krathwohl, 2001) provides a framework that is used to describe the relationship between students' questions and their conceptual and procedural knowledge. Following the levels of Bloom's Taxonomy, we can describe student questions as indicative of lower-order and higher-order thinking. Lower-order questions tend to be closed and consist of recall and comprehension-based knowledge, in line with procedural knowledge and understanding (Zepeda, 2008). In contrast, higher-order questions tend to align with conceptual knowledge (Hopper, 2006) and be more open and critical, involving analysis and evaluation of concepts and ideas (Hopper, 2006; Khan & Inamullah, 2011). Higher-order questions usually refer to those that require in-depth thinking, analysis, and synthesis of information. Educational research suggests that when using higher-order questions, elementary school students perform better on multiple-choice and essay portions of class tests (Barnett & Francis, 2012), and improve critical thinking skills (Achmad & Utami, 2023). Alharbi and colleagues (2022) believed that an online collaborative learning environment can significantly improve the grades of female students on tests about higher-order thinking skills;

The study of Yuliati and Lestari (2018) concluded that students' ability to answer higher-order questions improves with age. Renaud and Murray (2006) claimed that higher-order questions are related to gains in students' critical thinking skills and can be a valid process indicator.

Collaborative learning has proven to be an effective way to promote high-level thinking and higher-order questions (Earle-Randell, 2023; Tsan, 2019). It provides students with opportunities to develop significantly stronger problem-solving skills than they would individually (Fawcett, 2011), and research has established that supporting this discourse through collaborative learning has a positive impact on student's critical thinking skills (Warsah, 2021). Collaborative learning involves two or more learners working together on a shared learning goal through information sharing and negotiation (Dillenbourg, 1999; Roschelle & Teasley, 1995), and as a form of collaborative learning, pair programming, has been particularly effective in K-12 Computer Science (CS) Education and has been demonstrated to positively impact problem-solving skills and CS knowledge (Wei et al., 2021). There is a growing body of knowledge on the use of collaborative programming in K-12 classrooms (Earle-Randell et al., 2024; Zhong et al., 2016) but understanding at a more granular level the collaborative behaviors that emerge during pair programming activities and how we can support learners during these tasks is still an open question. Utilizing natural language processing has been a successful strategy for researchers to model the collaborative discourse between learners (Earle-Randell, 2023), but looking deeply into the higher-order questions that students ask during these collaborative problem-solving tasks could provide valuable insight into the behaviors that drive collaboration between K-12 learners.

3. Dataset

In this paper, we utilized a dataset that was collected as part of a larger project to investigate collaborative CS learning with virtual agents for upper elementary school children and has previously been published by Earle-Randell et al. (2023; 2024) and Ma et al. (2023). The dataset we analyzed consists of video and audio recordings of 44 fourth-grade learners in an elementary school in the southeastern United States who provided assent and parental consent. This study was conducted in a block-based coding environment called FLECKS (Zakaria et al., 2021), where student dyads used pair programming to collaborate on a series of coding activities in which they practiced fundamental CS concepts such as variables, conditionals, and loops. FLECKS is built upon a block-based coding environment called NetsBlox (NetsBlox, 2024), and it was designed to include two pedagogical virtual agents designed to foster good collaborative learning practices between the learner dyads by modeling positive collaborative behavior through brief vignettes that sometimes directly address the learners. The virtual agents remained on the screen throughout the session.

Once the data was collected, researchers transcribed each session. Our corpus contains 35 sessions and 9,996 utterances, where each utterance represents an uninterrupted chain of language spoken by an individual. Researchers modified Zakaria et al.'s (2021) dialogue act taxonomy to isolate "exploratory talk" dialogue and highlight question-asking behaviors, labeling each utterance in the dataset. The labels included two types of question asking, “Question-higher order” and “Question-other,” in addition to eleven other coding labels detailed in Table 1. In this paper, we focus on higher-order questions. This dialogue act taxonomy was applied by two annotators who were familiar with the context of the study. They independently applied labels to an overlapping 20% of the data, reaching a Cohen's Kappa score of 0.816, indicating a strong agreement. They then proceeded to divide and label the remaining data independently.

Table 1. *Dialogue Act Taxonomy Used in this Analysis*

Label	Frequency	Description
Agreement / Acknowledgement	1107	Agreement on a decision or opinion
Antagonistic Action	149	Actions that cause tension between the dyad
Confusion / Help-Seeking	417	Learner directly or indirectly seeks help
Directed at Agent	99	Something said directly to the agent
Directive	823	Telling their partner to do something
Disagreement / Negative feedback	715	Disagreement on a decision or opinion
Disagreement with Justification	34	Disagreement, but provides reasoning
Other	1890	Something not covered by other labels
Question - Higher-order	146	Asking WHY or challenging an idea
Question - other	1066	Asking anything other than a why question
Social	1081	Social/off-topic dialogue
Self-Explanation / Justification	1344	Explaining their thoughts or their steps
Suggestion / Alternative Idea	1114	Any idea when directly talking to their partner

4. Methodology

To address our research questions, we conducted two analyses in sequence using the labeled dialogue. First, we built a higher-order question classification model using TF-IDF (Term Frequency Inverse Document Frequency) and Sentence-BERT as feature extraction methods, applied separately across five algorithms: Logistic Regression, Decision Tree, Gaussian Naive Bayes, K-Nearest Neighbors, and XGBoost. We then evaluated the model performance by calculating AUC, Accuracy, F1, and Recall scores to determine which model performed better. In addition, we calculated a confusion matrix for the highest-performing models to further examine the most common labels. In the second analysis, we conducted a Hierarchical Clustering Analysis by clustering utterances labeled higher-order questions along with other dialogue acts identified through the Confusion Matrix to identify different types of higher-order questions and similar patterns within other dialogue acts. We will break down the methods into subsections in the following discussion.

To ensure that our text data could be used to fit machine learning models, we first processed it through text vectorization. We utilized two methods, TF-IDF and Sentence-BERT, to extract features and compare their effectiveness across different models.

TF-IDF is a text feature extraction method that considers the frequency of terms in the document and in the entire corpus and assigns corresponding weights. TF-IDF has become a popular method of creating features that describe documents, or dialogue acts in our context, based on the importance of the words contained within (O'Keefe & Koprinska, 2009); the importance of words is represented as a weight based on the frequency of that word within each document as compared to its appearance in all documents. Conversely, Sentence-BERT is a text representation model that is designed to represent semantic meaning by embedding strings of words (e.g. sentences) into a high-dimensional feature vector (Reimers & Gurevych, 2019). It uses a Siamese network structure to learn these sentence-level embeddings based on semantic similarity. Beyond this, pre-trained Sentence-BERT models have been made

available through a repository known as Huggingface (Huggingface, n.d.). In this study, we specifically use the “paraphgrase-MiniLM-L6-v2” pre-trained Sentence-BERT model.

Both of these featurization methods offer a different representation of language and, in conjunction with a prediction model, can help provide insights into the relationship between language patterns and the labels of higher-order questions. TF-IDF provides a measure of word importance while ignoring synonym-based relationships between those words, placing a larger emphasis on specific vocabulary. Sentence-BERT, conversely, places lower emphasis on specific word choices and instead captures semantic meaning. In comparing these two featurization methods, we can determine whether certain keywords (through TF-IDF) are indicative of higher-order questions, or if instead the semantic meaning of different, but similar, phrases (through Sentence-BERT) are more predictive of students’ higher-order questions.

To address our first research question, we employed 10-fold stratified cross-validation to evaluate the performance of five distinct types of classification model: Logistic Regression, Decision Tree, Gaussian Naive Bayes, K-Nearest Neighbors, and XGBoost. We selected a diverse array of algorithms due to the uncertainty surrounding model performance in this context. For comparison, we utilized two sets of features derived from TF-IDF and Sentence-BERT embeddings, respectively. Considering that our data has an imbalanced sample size, we compared the models based on commonly used machine learning model evaluation indicators (AUC, Kappa, Recall, and Accuracy) to examine model performance comprehensively. We also optimized the balance between the true positive rate and the false positive rate by leveraging the ROC curve to find the optimal threshold for classifying observations which may otherwise affect metrics of classification due to unbalanced labels such as Kappa, Recall, and Accuracy (c.f. Bosch & Paquette, 2018); the AUC metric is not sensitive to the choice of rounding threshold and is therefore unaffected by this optimization step. This optimization is calculated for each fold of each model based on the ROC curve produced by the respective training set.

To understand how other dialogue acts may confuse our higher-order question model, we calculated the Confusion Matrix for the two highest-performing models evaluated for RQ1. While the model only predicted higher-order Questions (1) and all other labels (0) as a binary prediction task, we examined the distribution of predictions across the other labels after applying the model and examined the labels for which there were disproportionate false-positive predictions.

To answer our second research question, we conducted a hierarchical clustering analysis to group utterances according to their similarity and identify clusters of higher-order questions. In our work, we used an agglomerative strategy with a silhouette score to find the optimal number of clusters (cf. Roux, 2018). To further understand how utterances from other labels are grouped and distributed with higher-order questions, we selected dialogue labels based on the confusion matrix results from the prior analysis, which indicated higher false positive rates among these labels, and used the clusters to identify different profiles of higher-order questions based on the proportion of utterances belonging to each grouping. To compare clustering and model performance, we identified the frequency of utterances predicted as higher-order questions within each cluster. In addition, we also identified the most frequent bigrams that occurred in each cluster.

5. Results

Logistic Regression and XGBoost using TF-IDF demonstrated better performance than the other models, as seen in Table 2, when accounting for all metrics. The Logistic Regression model exhibited the highest AUC score (0.97) and an optimized recall of 0.92. In contrast, XGBoost scored higher on Kappa values (0.4) but had a significantly lower Recall (0.52), highlighting the subtle differences in model performance, particularly on false positive rates.

Table 2. *Evaluation of Classifier Performance for TF-IDF and Sentence-Bert Datasets at Group Level*

	Model	AUC	Kappa	Optimized Kappa	Accuracy	Optimized Accuracy	Recall	Optimized Recall
TF-IDF	LR	0.97	0.23	0.38	0.99	0.96	0.16	0.92
	DT	0.74	0.43	0.44	0.98	0.98	0.47	0.48
	GNB	0.59	0.02	0.02	0.69	0.69	0.48	0.48
	KN	0.76	0.29	0.32	0.99	0.97	0.22	0.53
	XG	0.98	0.4	0.48	0.98	0.98	0.38	0.52
Sentence-BERT	LR	0.96	0.44	0.35	0.99	0.96	0.38	0.73
	DT	0.57	0.11	0.11	0.97	0.97	0.15	0.15
	GNB	0.92	0.18	0.13	0.92	0.88	0.7	0.77
	KN	0.77	0.22	0.27	0.99	0.96	0.14	0.56
	XG	0.94	0.19	0.22	0.99	0.99	0.13	0.15

Note. LR=Logistic Regression; DT=Decision Tree; GNB=Gaussian Naive Bayes; KN= K-Nearest Neighbors; XG=XGBoost

Among all dialogue acts, we discovered that the four dialogue acts most confusing to the higher-order question models were 'Social', 'Question - other', 'Directed at Agent', and 'Confusion/Help Seeking' (see Table 3). This misclassification implies similar language patterns underlying these interaction categories.

Table 3. *Accuracy of Logistic Regression and XGBoost Classifier for Different Labels with TF-IDF Features at Group Level*

		Logistic Reg. with TF-IDF		XGBoost with TF-IDF	
Label	N Total	% Correct	% Incorrect	% Correct	% Incorrect
Agreement / Acknowledgement	1107	98.83	1.17	99.82	0.18
Antagonistic action	149	95.3	4.7	100	0
Confusion / Help-Seeking	417	92.09	**7.91**	100	0
Directed at Agent	99	92.93	**7.07**	97.98	**2.02**
Directive	823	99.39	0.61	100	0
Disagreement / Negative feedback	715	98.18	1.82	99.86	0.14
Disagreement with Justification	45	97.78	2.22	100	0
Other	1890	97.46	2.54	99.63	0.37
Question - Higher-order	146	93.15	**6.85**	51.37	**48.63**
Question - other	1066	88.09	**11.91**	96.81	**3.19**
Social	1081	94.17	**5.83**	97.59	**2.41**

Self-explanation / Justification	1344	96.88	3.12	99.7	0.3
Suggestion / Alternative Idea	1114	96.77	3.23	99.28	0.72

To take a nuanced look at model performance, our clusters include utterances labeled as higher-order Questions, Other Questions, Confusion/Help Seeking, Directed at Agent, and Social. They were selected based on the Confusion Matrix results, which showed higher false positive rates among all labels. In total, we identified 40 clusters of utterances, of which four contained a notable proportion of higher-order questions (Figure 1). We printed out utterances (n = 2809) in each cluster with labels and predicted values from the Logistic Regression model.

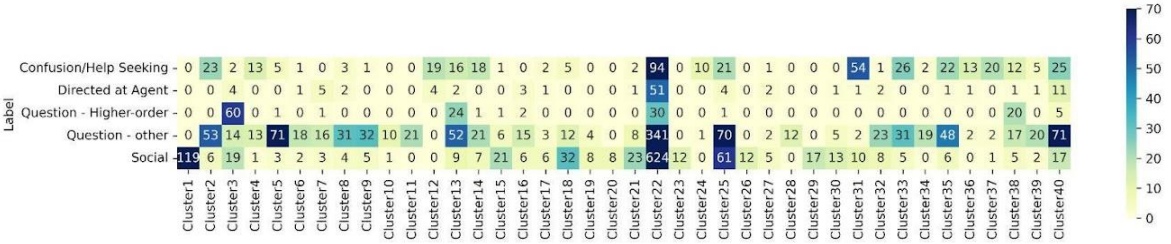

Figure 1. Heatmap of Cluster Distribution

While four clusters of higher-order questions emerged, utterances in other dialogue acts were also found to group with these questions. Clusters 3, 13, and 38 all have high false positive rates of 0.82, 0.44, and 1, respectively, suggesting similar language patterns underlying these interaction categories. More specifically, in Cluster 3, 'Other Questions' (14%) and 'Social' labels (19%) were misclassified the most. Clusters 13 and 22 exhibited significant misclassifications in 'Confusion' (16%, 8%) and 'Other-Question' (50%, 30%).

6. Discussion

To address the first research question, across all models, Logistic Regression and XGBoost using TF-IDF demonstrated the best performance. The high recall and low Kappa of Logistic Regression indicate that while it accurately captures most of these questions, overall reliability across all classifications might not be strong and it may struggle to distinguish between similar underlying patterns. In contrast, XGBoost's high Kappa and low recall indicate that while it has well-balanced decision boundaries that account for distinct categories, it might be too conservative in its classification criteria, missing predictions of some higher-order questions. Moving forward, a hybrid approach could leverage the advantages of both the Logistic Regression and XGBoost models to improve performance in classifying higher-order questions while maintaining well-balanced decision boundaries. This hybrid model could be applied in collaborative learning to provide real-time feedback on learners' higher-order questions and support effective problem-solving during a collaborative coding task.

Regarding the second research question, through close examination of higher-order questions in the four main clusters, we learned that the higher-order questions differed most notably in how they were phrased. The "why" questions were distinctive from "how" questions, and further distinctive from the "what if" questions. In discussions with their peers, students commonly asked "Why" and "How" questions, such as 'Why did you delete it?' and 'How do you make him go up?'. These examples show instances where students were trying to understand the reasoning and processes behind their partners' actions. According to Bloom's taxonomy (Anderson & Krathwohl, 2001), these questions engaged cognitive processes like 'understanding' and 'analyzing' as students sought to interpret their partner's behavior and make connections with explanations. Another form of higher-order question involved students aiming to understand the project by asking questions like 'Why did he stop moving?' and 'How do we make it walk in a square?' These questions engaged primarily in 'analysis' but also exemplified 'evaluation,' where the students focused on identifying problems in their code and examining the causes and effects. This aligns closely with conceptual knowledge, with

students displaying implicit knowledge of the principles behind the task and critically thinking about the problem (Hiebert & Lefevre, 1986; Hopper, 2006). The final form of higher-order question involves the exploration and generation of ideas with ‘What if’ or ‘How’ questions: ‘What if we clicked that?’; ‘What if we change this off and then change it to this?’ These questions engage the ‘creating’ level of Bloom’s Taxonomy by hypothesizing and proposing alternative solutions to the problem. This distinctly aligns with conceptual knowledge, indicating that students are building upon the procedural skills necessary to complete the task and synthesizing new ideas to solve the problem (Rittle-Johnson, et al., 2001).

Based on the results from the Confusion Matrix of Logistic Regression and XGBoost models and cluster analysis, the four dialogue acts that have comparable patterns and were frequently confused with the higher-order question model were 'Social', 'Question - other', 'Directed at Agent', and 'Confusion/Help Seeking' (Table 3). Many of these utterances appear to be questions directed at others that follow a similar structure to the other higher-order questions, emphasizing the importance of considering the context within the coding scheme. Other labels, particularly Confusion/Help-seeking, appeared to use some language similar to that of a higher-order question but these dialogue acts were expressed as statements rather than questions (Table 4). It is important to note that dialogue acts labeled as Confusion/Help-seeking often exhibit higher-order thinking skills, which could be important to consider when evaluating how students are collaboratively problem-solving. Research shows behaviors of confusion or help-seeking demonstrate students’ engagement and the use of adaptive learning strategies, signaling deep cognitive processing (D’Mello et al., 2014). Our findings on the interplay between higher-order questions and confusion/help-seeking further suggest that these behaviors reflect critical thinking through higher-order thinking evidence in the students’ utterances, which is also a signal of effective learning. Embedding more contextual understanding into the model could help disambiguate these patterns and establish the relationship between higher-order questions and confusion/help-seeking behaviors.

Table 4. *Examples of Dialogue Acts that Confused the Model*

Higher-Order Question	Confusion / Help-Seeking	Direct at Agent	Other-Question	Social
“Why is it going so slowly now?”	“I don't know how to do it”	“That's my question, why?”	“Repeat, how many times?”	“Can you hear me?”

7. Limitations and Future Work

While our models achieved a good performance in detecting higher-order questions, we acknowledge the limitations of this study. All of the models treat each utterance as an independent instance and may miss the nuances that context-based analysis could offer.

Future directions for this work include a context-driven approach to this classification task. Focusing on the semantics and surrounding dialogue of each utterance could improve the accuracy of the model, allowing for a clearer understanding of the higher-order questions students are asking. Additionally, the relationships between higher-order questions and other similarly structured utterances, particularly confusion and help-seeking, warrant further investigation. Exploring the nuances of these intertwined dialogue acts could provide valuable insights into the dynamics of effective collaboration and higher-order thinking. Investigating how higher-order questions interact with confusion and help-seeking behaviors can provide us with a richer understanding of how young learners navigate confusion through higher-order thinking and discourse moves. Future analysis could inform guidelines for productive collaboration, scaffolding when and how to overcome roadblocks with higher-order questions, and creating environments that foster engaging and effective collaborative problem-solving.

8. Conclusion

This study explores the underlying patterns of students' higher-order questions to unpack their learning processing during collaborative problem-solving. We first developed binary classification models using TF-IDF and XGBoost features across various algorithms to assess their effectiveness in predicting higher-order questions in problem-solving collaborative discourse during a pair programming task. Our models show Logistic Regression and XGBoost have a reliable performance with high accuracy scores, though we observed a trade-off between Kappa and Recall. To further assess the model performance, we created a Confusion Matrix and found that labels, Other Questions, Confusion/Help-seeking, Directed at Agent, and Social share similar language patterns, which may confuse the model due to the similar characteristics of utterances. To deepen our understanding of the scenarios involving higher-order questions and how other dialogue acts may interact with them, we employed a hierarchical cluster strategy and identified three types of higher-order questions: 1) students seeking to understand the reasoning and processes behind their partners' actions, 2) students aiming to understand and analyze the project by pinpointing the problems, 3) students generating ideas by proposing the 'What if' or 'How' question, which stimulated deeper cognitive processing. These distinct types of higher-order questions exemplify students utilizing their conceptual knowledge to think about a problem deeply and critically (Hopper, 2006; Rittle-Johnson et al., 2001) and align with categories in Bloom's Taxonomy.

The Confusion/Help-Seeking, Directed at Agent, and Social dialogue acts that share similar linguistic patterns to higher-order questions highlight how complex and dynamic dialogue with questions can be. It is important to recognize the intentions and context behind these questions, establishing the relationships between higher-order questions and other, similar utterances like confusion and help-seeking. This can help us further define the dialogue that embodies higher-order thinking and inform effective feedback in an authentic learning environment. Further understanding the relationship between confusion and higher-order questions can help us encourage students to collaborate and overcome challenges effectively.

The insight gained from analyzing higher-order question patterns during collaborative problem-solving may be extrapolated to identify different types of higher-order questions across other contexts, guiding teachers in promoting higher-order questions and supporting the development of adaptive educational technologies that analyze discourse and aid learners.

Acknowledgments

This research was supported by the National Science Foundation through grants IIS-2331379 and DRL-1721160 and jointly by the National Science Foundation and the Institute of Education Sciences under grant DRL-2229612. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the U.S. Department of Education.

References

Achmad, W. K. S., & Utami, U. (2023). High-order questions improve students' critical thinking skills in elementary schools. *International Journal of Elementary Education*, *7*(2), Article 2.

Alharbi, S. M., Elfeky, A. I., & Ahmed, E. S. (2022). The effect of e-collaborative learning environment on development of critical thinking and higher order thinking skills. *Journal of Positive School Psychology*, 6848–6854.

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives*. Longman.

Barnett, J. E., & Francis, A. L. (2012). Using higher order thinking questions to foster critical thinking: A classroom study. *Educational Psychology*, *32*(2), 201–211.

Boldero, J., & Fallon, B. (1995). Adolescent help-seeking: What do they get help for and from whom? *Journal of Adolescence*, *18*(2), 193–209.

Bosch, N., & Paquette, L. (2018). Metrics for discrete student models: Chance levels, comparisons, and use cases. Journal of Learning Analytics, 5(2), 86–104.

Bustami, Y., Syafruddin, D., & Afriani, R. (2018). The implementation of contextual learning to enhance biology students' critical thinking skills. *Jurnal Pendidikan IPA Indonesia*, *7*(4), Article 4.

Dillenbourg, P. (1999). *What do you mean by collaborative learning?* (p. 1). Oxford: Elsevier.

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145–157.

D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, *29*, 153–170.

Earle-Randell, T. V., Wiggins, J. B., Ma, Y., Celepkolu, M., Bounajim, D., Gao, Z., Ruiz, J. M., Boyer, K. E., Israel, M., Lynch, C. F., & Wiebe, E. (2024). The impact of near-peer virtual agents on computer science attitudes and collaborative dialogue. *International Journal of Child-Computer Interaction*, *40*, 100646.

Earle-Randell, T. V., Wiggins, J. B., Ruiz, J. M., Celepkolu, M., Boyer, K. E., Lynch, C. F., Israel, M., & Wiebe, E. (2023). Confusion, conflict, consensus: Modeling dialogue processes during collaborative learning with hidden markov models. *Artificial Intelligence in Education: 24th International Conference Proceedings*, 615–626.

Fawcett, L. (2002). The effect of peer collaboration on children's problem solving ability. *Theses : Honours*. https://ro.ecu.edu.au/theses_hons/921

Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In *Conceptual and procedural knowledge: The case of mathematics* (pp. 1–27). Lawrence Erlbaum Associates, Inc.

Hopper, C. H. (2006). *Practicing college learning strategies*. Houghton Mifflin Company.

Huggingface (n.d.). paraphrase-MiniLM-L6-v2. Retrieved from https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2

Karabenick, S. A., & Knapp, J. R. (1991). Relationship of academic help seeking to the use of learning strategies and other instrumental achievement behavior in college students. *Journal of Educational Psychology*, *83*(2), 221–230.

Khan, W., & Inamullah, H. (2011). A study of lower-order and higher-order questions at secondary level. *Asian Social Science*, *7*(9), Article 9.

Ma, Y., Celepkolu, M., Boyer, K. E., Lynch, C. F., Wiebe, E., & Israel, M. (2023). How noisy is too noisy? The impact of data noise on multimodal recognition of confusion and conflict during collaborative learning. *International Conference on Multimodal interaction,* 326-335.

Netsblox—Innovative learning. (n.d.). Retrieved April 29, 2024, from https://netsblox.org/

O'Keefe, T., & Koprinska, I. (2009). *Feature selection and weighting methods in sentiment analysis*. 67–74.

Puustinen, M. (1998). Help-seeking behavior in a problem-solving situation: Development of self-regulation. *European Journal of Psychology of Education*, *13*, 271–282.

Qayyum, A. (2018). Student help-seeking attitudes and behaviors in a digital era. *International Journal of Educational Technology in Higher Education*, *15*(1), 17.

Reimers, N., & Gurevych, I. (2019). *Sentence-bert: Sentence embeddings using siamese bert-networks* (arXiv:1908.10084). arXiv.

Renaud, R. D., & Murray, H. G. (2007). The validity of higher-order questions as a process indicator of educational quality. *Research in Higher Education*, *48*(3), 319–351.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology*, *93*(2), 346–362.

Roux, M. (2018). A comparative study of divisive and agglomerative hierarchical clustering algorithms. *Journal of Classification*, *35*, 345-366.

Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In C. O'Malley (Ed.), *Computer Supported Collaborative Learning* (pp. 69–97). Springer.

Tsan, J., Vandenberg, J., Fu, X., Wilkinson, J., Boulden, D., Boyer, K. E., Lynch, C., & Wiebe, E. (2019). *Conflicts and collaboration: A study of upper elementary students solving computer science problems*. https://repository.isls.org//handle/1/4521

Warsah, I., Morganna, R., Uyun, M., Afandi, M., & Hamengkubuwono, H. (2021). The impact of collaborative learning on learners' critical thinking skills. *International Journal of Instruction*, *14*(2), Article 2.

Wei, X., Lin, L., Meng, N., Tan, W., Kong, S.-C., & Kinshuk. (2021). The effectiveness of partial pair programming on elementary school students' computational thinking skills and self-efficacy. *Computers & Education*, *160*, 104023.

Williams, L., Wiebe, E., Yang, K., Ferzli, M., & Miller, C. (2002). In support of pair programming in the introductory computer science course. *Computer Science Education*, *12*.

Yuliati, S., & Lestari, I. (2018). Higher-order thinking skills (hots) analysis of students in solving hots question in higher education. *Perspektif Ilmu Pendidikan*, *32*, 181–188.
Zakaria, Z., Vandenberg, J., Tsan, J., Boulden, D. C., Lynch, C. F., Boyer, K. E., & Wiebe, E. N. (2022). Two-computer pair programming: Exploring a feedback intervention to improve collaborative talk in elementary students. *Computer Science Education*, *32*(1), 3–29.
Zepeda, S. J. (2008). *The instructional leader's guide to informal classroom observations*. Routledge.
Zhong, B., Wang, Q., & Chen, J. (2016). The impact of social factors on pair programming in a primary school. *Computers in Human Behavior*, *64*, 423–431.