Collaborative Game-based Learning Analytics: Predicting Learning Outcomes from Game-based Collaborative Problem Solving Behaviors

Halim Acosta
North Carolina State University
Raleigh, USA
hacosta@ncsu.edu

Daeun Hong
Indiana University
Bloomington, USA
dh37@iu.edu

Seung Lee
North Carolina State University
Raleigh, USA
sylee@ncsu.edu

Wookhee Min
North Carolina State University
Raleigh, USA
wmin@ncsu.edu

Bradford Mott
North Carolina State University
Raleigh, USA
bwmott@ncsu.edu

Cindy Hmelo-Silver
Indiana University
Bloomington, USA
chmelosi@indiana.edu

James Lester
North Carolina State University
Raleigh, USA
lester@csc.ncsu.edu

Abstract

Skills in collaborative problem solving (CPS) are essential for the 21st century, enabling students to solve complex problems effectively. As the demand for these skills rises, understanding their development and manifestation becomes increasingly important. To address this need, we present a data-driven framework that identifies behavioral patterns associated with CPS practices and can assess students' learning outcomes. It provides explainable insights into the relationship between students' behaviors and learning performance. We employ embedding and clustering techniques to categorize similar trace logs and apply Latent Dirichlet allocation to generate meaningful descriptors. To capture the temporal evolution of student behaviors, we introduce a graph-based representation of transitions between behavior patterns extracted using constraint-based pattern mining. We map behavioral patterns to a CPS ontology by analyzing how action sequences correspond to specific CPS practices. Analysis of semi-structured trace log data from 61 middle school students engaged in collaborative game-based learning reveals that the extracted behavioral patterns significantly predict student learning gains using generalized additive models. Our analysis identifies patterns that provide insights into the relationship between student use of CPS practices and learning outcomes.

CCS Concepts

• Computing methodologies → Machine learning; **• Applied computing → Collaborative learning**; **Interactive learning environments**.

LAK 2025, Dublin, Ireland

ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706522

Keywords

Collaborative Game-Based Learning, Collaborative Problem Solving, Trace Log Analysis, Explainable Machine Learning

ACM Reference Format:
Halim Acosta, Daeun Hong, Seung Lee, Wookhee Min, Bradford Mott, Cindy Hmelo-Silver, and James Lester. 2025. Collaborative Game-based Learning Analytics: Predicting Learning Outcomes from Game-based Collaborative Problem Solving Behaviors. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3706468.3706522

1 Introduction

Collaborative game-based learning environments foster the development of 21st century skills while providing researchers with data to analyze collaborative dynamics. These environments generate behavior-rich interaction data that can reveal student learning processes and outcomes, holding significant potential to offer adaptive scaffolding at individual, group, and whole class levels within digital learning environments [25, 28]. While identifying salient patterns from student problem-solving behaviors is essential for assessing knowledge and skill acquisition in game-based learning [42], a key challenge lies in interpreting collaborative learning-based raw trace log data within educational theories. Effectively analyzing the semi-structured nature of the data is crucial for understanding students' behaviors during collaborative problem solving (CPS) and assessing their learning outcomes.

Frameworks, such as OECD PISA and the CPS ontology, assess collaborative problem-solving skills along social and cognitive dimensions [2, 18, 30]. However, traditional assessments (e.g., exams, surveys) only provide static snapshots of learning, failing to capture real-time learning processes [33, 38]. These methods typically overlook the evolving nature of collaboration, making it difficult to assess how students respond and adapt during the learning process. Additionally, while digital learning environments generate rich interaction data, there is limited research investigating how well the collected behavior data map to important learning constructs [35, 43]. Further research is needed to explore how behavioral patterns extracted from process data can reliably represent collaborative dynamics and inform relevant interventions.

There are many existing methods for extracting sequential patterns from interaction data, such as the Generalized Sequential Patterns (GSP) algorithm [36], constrained Sequential Pattern Discovery using Equivalence classes (cSPADE) [44], Prefix Span [32], and SPAM [3]. Other pattern analysis methods exist, such as lag sequential analysis (LSA), process mining, and epistemic network analysis [4, 6, 31]. The addition of constraints to sequential pattern mining (SPM) methods incorporates filtering techniques for unreliable patterns, making them well suited for incorporating domain knowledge into the pattern mining process [45].

Our work uses clustering and constraint-based sequential pattern mining (CSPM) to extract and analyze pedagogically relevant behavioral patterns by mapping in-game behaviors to the CPS ontology proposed by Andrews-Todd and Forsyth [2]. Our approach interprets the extracted patterns through pedagogical theory and reveals the relationship between CPS behaviors and learning outcomes via predictive modeling. We apply the Seq2Pat CSPM algorithm [41], which leverages multi-valued decision diagrams [19], to student game trace logs, uncovering CPS behaviors while filtering out irrelevant actions. Our research addresses the following questions:

- **RQ1:** Can CPS behaviors in a collaborative game-based learning environment be extracted from raw trace data using CSPM?
- **RQ2:** How well do these behavioral patterns predict learning outcomes in a collaborative game-based learning environment?
- **RQ3:** How do static artifacts compare to temporal artifacts in their relationship with student learning outcomes in collaborative game-based learning activities?

Our findings suggest that static and temporal artifacts effectively predict student performance, with static-based features providing the highest predictive performance averaged across all tests. Additionally, we reveal linear and non-linear interactions between behaviors and learning outcomes, offering insights for personalized learning and scaffolding.

2 Related Work

2.1 Collaborative Problem Solving

Collaborative problem solving (CPS) involves two or more individuals integrating their knowledge, skills, and efforts to solve a problem [7, 30]. As a critical 21st century skill, CPS has been extensively studied [2, 18, 21, 30]. CPS integrates cognitive and social dimensions, with the cognitive dimension focusing on problem-solving and the social dimension on collaboration [18, 37]. For example, the OECD's PISA 2015 framework outlines four cognitive aspects—exploring, representing, planning, and monitoring—and three social aspects—shared understanding, action-taking, and team organization [30]. The ATC21S framework emphasizes subskills like participation, perspective taking, and social regulation, as well as task regulation and knowledge building on the cognitive side [18].

Andrews-Todd and Forsyth [2] conceptualized a CPS model as an ontology, building on a literature review of CPS within computer-supported collaborative learning (CSCL) environments and related fields. They argued that CPS competencies manifest not only through actions but also through discourse practices, such as discussions. In their theoretical model, the social dimension, which relates to collaborative processes, includes 1) maintaining communication, 2) sharing information, 3) establishing shared understanding, and 4) negotiating. Maintaining communication involves practices that are socially oriented rather than directly related to hands-on tasks, such as using chat emojis, greeting and complimenting teammates, and apologizing. Sharing information pertains to providing task-related or resource information and ideas to others. Establishing shared understanding refers to integrating and consolidating agreed-upon knowledge from others (e.g., "That makes sense"). Negotiation involves identifying and resolving conflicts (e.g., "Does anyone disagree?" and "Okay, I'm convinced"). The cognitive dimension, associated with problem-solving processes, includes 1) exploring and understanding, 2) representing and formulating, 3) planning, 4) executing, and 5) monitoring. Exploring and understanding refers to activities aimed at making sense of the problem, relevant content knowledge, and the task at hand. Representing and formulating involve actions and communications that demonstrate knowledge. Planning corresponds to developing strategies for problem solving, while executing refers to carrying out these plans. Lastly, monitoring involves actions and communications to track progress towards the goal and assess the status of teammates.

Previous studies have applied this model in science-based, computer-mediated learning environments using process data, like time-stamped log files, to measure complex skills such as collaboration [2, 9, 37], moving beyond traditional assessments [5, 16, 20, 39]. These studies measured and analyzed students' CPS competencies by using text-mediated communication and process data as evidence of CPS practices, mapping this evidence to each dimension of the CPS framework. However, due to the inherent complexity of CPS, assessing students' performance remains challenging, particularly within complex environments like CSCL classrooms. Furthermore, detecting and diagnosing student engagement at lower levels of CPS practices—rather than focusing solely on higher-level skills like the social and cognitive dimensions—has remained elusive [2, 39]. In response, the current study adopts an interdisciplinary approach to more effectively examine CPS competencies within a middle school science context in a CSCL environment.

2.2 Student Behavior Analysis with Game Trace Logs

The increasing availability of learner interaction data has enabled researchers to mine and analyze log data for insights into user interactions [15]. Researchers use behavioral logs from digital learning environments to analyze group behavior [17], predict initial student knowledge [13], discover problem-solving characteristics in puzzle games [40], and uncover contextual behavior patterns through causal relationships in event logs [1], extending even to non-educational settings [26].

Sequential mining techniques with constraints have proven effective for extracting behavioral patterns from game-based learning environments [22]. Researchers identified efficient strategies during different gameplay phases by applying hidden Markov models (HMMs) and a constrained Prefix Span algorithm, explaining how students advance through problem-solving processes. A review of sequential pattern mining (SPM) for educational data reveals its effectiveness in studying self-regulated learning and its advantages over simple event counts or durations when combined with statistical inference [45]. SPM additionally provides a flexible approach to capturing local learning patterns [45]. Lu et al. [23] used game telemetry data to build predictive models linking learning outcomes with log-generated information. Their interpretable models identified in-game behaviors associated with expertise and learning achievements.

Beyond game-based environments, sequence mining techniques have been applied to broader educational contexts. For example, researchers analyzed sequence data from Moodle interactions to predict student learning styles using clustering and machine learning models, achieving high accuracy and Kappa scores [12]. Another approach proposed a framework based on Rational Action Theory (TRA) and the Technology Acceptance Model (TAM). Dou et al.'s framework aggregates behavioral data and identifies features indicative of students' behavior, showcasing the utility of process data for deeper insights into student's behavioral intelligence [10]. Recent advances in pattern discovery algorithms, such as the Seq2Pat algorithm, have further expanded the potential of sequence mining. Ghosh et al. [14] demonstrated that dichotomic pattern mining (DPM) and constraint-based sequential pattern mining (CSPM) can effectively represent digital behavior for predictive modeling. CSPM offers strong integration for digital behavior analysis, while DPM enhances performance in complex models through feature augmentation.

Our research applies CSPM to student log data from a collaborative game-based learning environment. We use CSPM to generate static and temporal data representations as features for explainable statistical modeling techniques. We aim to explore the relationship between CPS-related behavioral patterns and student learning outcomes to create a robust framework for identifying pedagogically relevant behaviors from raw game trace logs.

3 EcoJourneys Learning Environment

EcoJourneys is a collaborative game-based learning environment designed to enhance middle school students' collaborative problem-solving skills and understanding of life science topics (Figure 1). In the game, learners form teams of three to four to investigate the cause of an illness spreading among the tilapia fish on a local farm [29, 34]. The game includes a tutorial and three quests, each structured around a collaborative inquiry framework that consists of three phases: 1) Talk & Investigate, 2) Deduce, and 3) Explain. At the start of each quest, students gather and investigate details about water quality and aquatic ecosystems by taking notes, interacting with non-player characters (NPCs), and viewing videos while exploring the game's environment during the Talk & Investigate phase. Following this, students proceed to the Deduce and Explain phases, which primarily involve applying their CPS skills.

Figure 1: EcoJourneys learning environment.

As illustrated in Figure 2, during the Deduce phase, students collaboratively determine answers to multiple-choice questions that provide scaffolding for interpreting the data they collected, ensuring that learners understand the significance of the data they will need for the final phase. They share relevant information and ideas, negotiate differences in opinions, and address any knowledge gaps by revisiting in-game learning resources such as their notebook or videos. After deliberation, they reach a consensus on a single answer before submitting their responses. The game provides feedback on the accuracy of their answer after each submission. If their response is incorrect, they must revisit the task and discuss it further with their group members. Students collaboratively answer a constructed response question at the end of the Deduce phase. Throughout the game, students primarily communicate via an in-game chat feature. In the Explain phase, students use a structured whiteboard to evaluate whether the information they have gathered (e.g., a note) supports a particular claim. They argue for or against the claim by placing relevant notes as evidence in the appropriate column on the board and explaining their reasoning to each other. In this study, we focus on the Deduce tasks across the three quests to examine patterns of students' in-game actions during CPS activities.

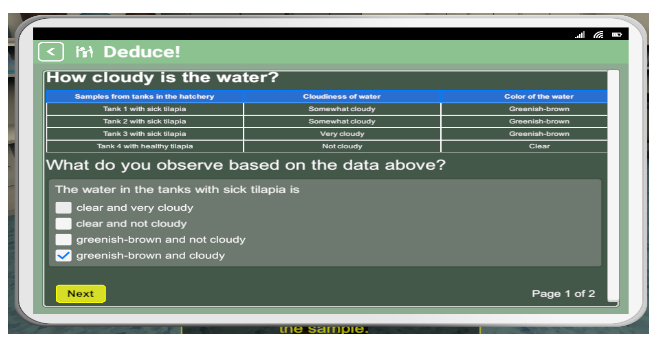

Figure 2: Example Deduce activity within EcoJourneys.

3.1 Dataset

We conducted an IRB-approved study with 76 middle school students in sixth through eighth grade (ages 11 to 14) using EcoJourneys. We collected data during classroom implementations of the game from Fall 2022 through Spring 2024, with students interacting with EcoJourneys for an average of one hour per gameplay session, with approximately 10 sessions taking place over a span of two weeks for each implementation. After removing students with missing post-test information, we used data from 61 students in the analysis presented in this paper. Throughout students' interactions with EcoJourneys, we collected detailed game trace data that logs their in-game actions. We use these game trace logs in our analysis.

4 Methodology

4.1 Data Preprocessing

In EcoJourneys, we record students' in-game actions in semi-structured trace log files. These logs capture interactions such as movement between locations, communication with NPCs, and group discussions via in-game chat, along with other miscellaneous data such as user-interface interactions and workbook engagement. The logs are stored in CSV format, with each event containing a timestamp and various context name-value pairs, leading to mixed data types that complicate sequential pattern mining analysis. Our behavioral pattern extraction framework is illustrated in Figure 3.

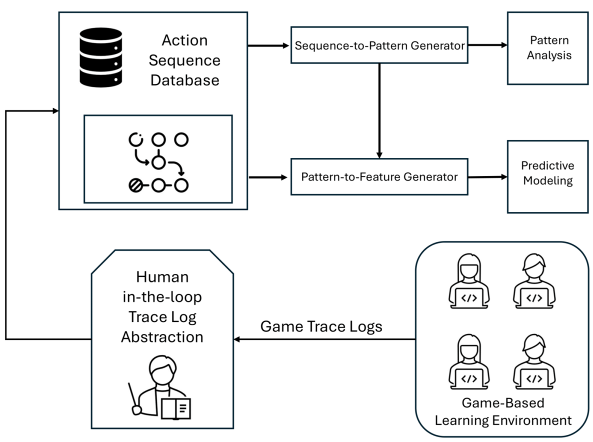

Figure 3: Behavioral pattern extraction framework.

Each trace event is annotated with a high-level action intent by preprocessing the context name-value pairs into string-based representations, which are tokenized, lower-cased, and cleaned of special characters. A Word2Vec model, trained on prior EcoJourneys logs, captures contextual relationships between events, chosen over transformer models due to lower computational cost and a smaller training corpus. Word2Vec also captures context better than lighter methods like TF-IDF. We generate an event embedding by averaging the input sequence and use K-means clustering with silhouette score to identify ten clusters. We refine these by selecting the most similar events based on cosine similarity using a threshold of 0.8, which was identified as the most effective in our preliminary analysis. Latent Dirichlet allocation then extracts topics from each cluster, enabling researchers to provide final descriptive cluster names. This results in a sequence of cluster labels corresponding to each student's actions. We apply CSPM (detailed in the next section) to these annotated sequences to extract behavioral patterns linked to CPS practices.

Our analysis focuses on student behaviors during collaborative interactions to understand the relationship between these behaviors and learning gains. Students alternate between independent exploration and collaboration at designated intervals in EcoJourneys during Deduce phases in the tutorial and quests. For this study, we isolate trace log data to cover events occurring specifically during collaborative interactions, as these are most relevant to CPS practices.

4.2 Constraint-Based Sequential Pattern Mining

In addressing our first research question, we utilize CSPM to extract behavioral patterns from raw trace log files. Each trace log file consists of a sequence of student in-game actions. By applying the clustering method described earlier, these events are generalized for pattern analysis. Using Seq2pat [45], we define constraints to refine the extracted patterns, an effective approach for educational data analysis. We apply a temporal constraint based on the time elapsed between consecutive trace events, ensuring patterns do not exceed 120 seconds to avoid disconnected events or noise from extended time frames. Additionally, a gap constraint limits patterns to a maximum of five repetitions of any event type, preventing the inclusion of looping behaviors and reducing spurious patterns. These parameters were chosen through a preliminary analysis that highlighted that too many or too few possible patterns hurt predictive models. Further exploration of the effects of constraints on the quality of extracted patterns is a promising direction for future work.

Seq2pat also allows control over the max span and minimum frequency to further restrict returned patterns. Given the frequent repetition of certain elements in trace logs, many patterns are pedagogically irrelevant or reflect routine game behaviors. To address this, we remove patterns occurring more than 100 times as part of our preprocessing to eliminate noise from the analysis. Preliminary analysis found that patterns that occur above the chosen threshold tend to represent routine system-level behaviors not having a strong pedagogical relevance.

4.3 Static & Graph-Based Data Representation

Seq2Pat extracts subsequences representing patterns from student action sequences. To utilize these patterns for predictive modeling, we generate static and graph-based representations of student behaviors. Static representations, which represent the presence or absence of an extracted behavioral pattern, consist of binary vectors indicating the presence or absence of patterns, yielding a high-dimensional matrix that we reduce by eliminating features with multicollinearity above 0.9. Initial investigations found that a smaller threshold removed very few patterns resulting in poor performance for predictive modeling. For the graph-based representation, reflecting students' transitions between various patterns, we create a transition matrix summarizing student behavior evolution. Directed graphs are constructed, where nodes represent patterns and edges denote transitions between them, with node values as transition counts and edge lengths based on intervening actions.

We compute graph attributes to enhance analysis: assortativity identifies if similar behaviors cluster, cluster coefficients show if certain patterns co-occur, and average shortest path length measures student transitions between patterns. Graph transitivity indicates common behavior sequences. The HITS algorithm identifies hubs (nodes with many outgoing links) and authorities (nodes with many incoming links), highlighting key patterns. Graph diameter and density provide insights into behavioral diversity and isolated pattern clusters, while degree centrality reveals focal behaviors. Both static and graph-based features are used for downstream modeling.

4.4 Predictive Modeling

To address our second research question, we analyzed the effectiveness of static and graph-based representations in predicting learning gains. High and low learning gains were defined relative to the median score. We explored explainable models like Logistic Regression, Elastic Net, Gradient Boosting, and Logistic Generalized Additive Models (GAM). Modeling student behaviors with GAMs provides a flexible approach for capturing non-linear relationships from high-dimensional data [27]. Logistic regression and Gradient Boosting have been shown to be effective in predicting educational outcomes while ensuring transparency [11, 24]. These models prioritize interpretability, aiding in the study of student behaviors and their relationship with learning outcomes. Given the small dataset, we used 10-fold stratified cross-validation with 5-fold nested cross-validation to optimize model hyperparameters. Model performance was evaluated using accuracy and F1-score.

For static features (high-dimensional binary vectors), we used Non-negative Matrix Factorization (NMF) for dimensionality reduction, as it offers a more interpretable parts-based representation compared to PCA, which may not effectively capture binary variable relationships well [8]. For graph-based features (floating point values), we used PCA to reduce dimensionality. We optimized models through 5-fold nested cross-validation. Logistic Regression hyperparameters included L1 or L2 penalties, Gradient Boosting parameters covered feature selection, tree depth, and number of estimators, and Elastic Net combined L1 and L2 penalties with stochastic gradient descent. Logistic GAMs were optimized for spline terms and smoothing parameters, controlling the bias-variance tradeoff. The best hyperparameters from nested cross-validation were used for the final model evaluation.

4.5 Explainability & Pattern Analysis

Our third research question investigates the relationship between static and graph-based representations and learning outcomes. We used the best-performing predictive model to explore this and analyzed feature contributions across cross-validation folds. We computed the Wald statistic and p-values for each fold, aggregating results by averaging the Wald statistic and combining p-values using Fisher's method. This approach allowed us to assess the overall significance of input features and their relationship with learning.

While these methods reveal the significance of principal components, we examined the loadings from our dimensionality reduction techniques (PCA and NMF) to understand the contribution of specific features. Loadings indicate the influence of features on principal components—PCA loadings represent the linear combination of features forming the principal components, while NMF loadings are the coefficients in the matrix that reconstruct the original input. We identified features with strong positive or negative relationships to principal components by examining loading values. Our analysis focused on the most significant components, highlighting the top three features contributing to predictive performance for deeper insight into their influence.

4.6 CPS Labeling Framework

After identifying significant CPS patterns from predictive modeling, the clusters within these patterns were mapped to one of the CPS codes from the CPS ontology [2], which served as the guiding theoretical framework (see Table 1). Clusters related to communication or interaction, such as 'Passive Communication Receipt' and 'Active Communication Initiation,' were classified under the broad 'social' dimension of the CPS framework rather than being assigned to more specific categories. This approach was used because accurately categorizing specific CPS practices (e.g., maintaining communication, sharing information, establishing shared understanding, and negotiating) requires discourse data to analyze the content of students' interactions [2, 37].

CPS practices within the cognitive dimension were identified through students' in-game behaviors captured in-game trace logs. For example, interactions involving in-game objects, resources like the notebook, and NPCs were categorized under 'Exploring and Understanding (EU).' Clusters such as 'Termination of Specific Interactions,' 'Workbook Engagement and Note-Taking,' and 'Tutorial Progression Engagement' were mapped to 'EU.' Similarly, clusters reflecting answer formulation and submission, like 'Interaction and Exploration Cessation' and 'Answer Submission and Validation', were aligned with 'Representing and Formulating (RF).' NPC dialogues after answer submission, which provided feedback on group performance, were categorized as 'Monitoring (M)' behaviors.

Some clusters, like 'Location-Based Activities and State Saving,' did not fit into any CPS category and were excluded. When clusters with the same label appeared consecutively, the label was assigned only once until a different code emerged. For instance, a pattern with 'Active Exploration and Interaction Initiation', 'Workbook Engagement and Note-Taking', and 'Answer Submission and Validation' would be encoded as an EU & RF pattern. Using these guidelines, CPS clusters were systematically mapped to relevant CPS ontology codes, providing a detailed understanding of students' CPS practices from game trace logs.

5 Results & Discussion

5.1 Predictive Modeling Results

For static features, Figure 4 presents the predictive performance of models averaged over 10-fold cross-validation. Evaluating data from different activities (Tutorial, Quests 1-3), we found that overall the logistic GAM model significantly outperforms other models. It achieved an average accuracy of 70.4% and an F1-score of 68.2%. In comparison, Gradient boosting and elastic net performed worse, with accuracies of 59.2% and 60.9%, respectively. The accuracy of the logistic GAM model was particularly good in Quests 1 and 3, with accuracies of 81.9% and 72.6%.

CPS Code	Description	Trace Log Cluster
Social	Communication through the in-game chat	Passive Communication Receipt
		Active Communication Initiation
Exploring & Understanding	Exploring the game environment by interacting with objects and resources	Termination of Specific Interactions
		Workbook Engagement and Note-Taking
		Active Exploration and Interaction Initiation
		Tutorial Progression Engagement
Representing & Formulating	Answering and submitting multiple-choice or constructed-response questions	Answer Submission and Validation
		Interaction and Exploration Cessation
Monitoring	Receiving feedback from NPC conversations	NPC Dialogue and Conversation Initiation

Table 1: Mapping of trace log clusters to CPS codes based on CPS ontology.

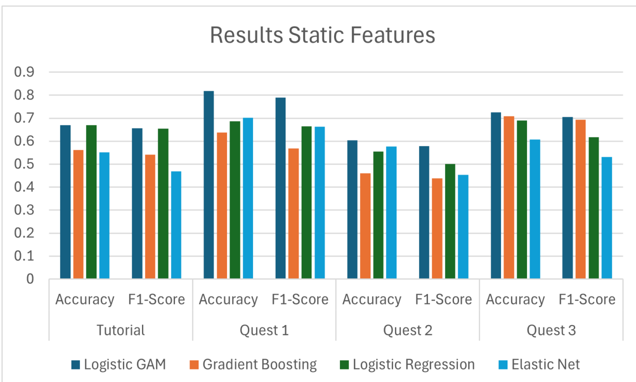

Figure 4: Predictive modeling results for static representations of behavioral patterns.

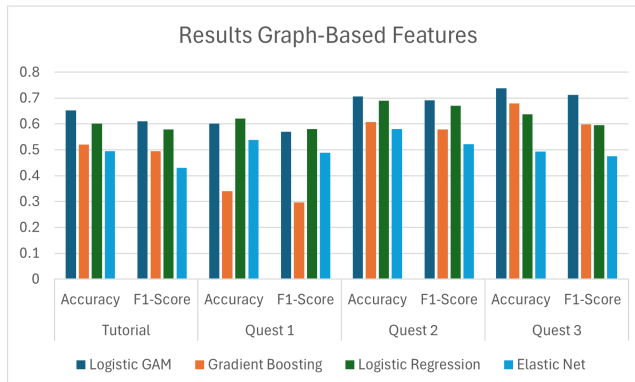

Figure 5: Predictive modeling results for graph-based representations of behavioral patterns.

For graph-based features, Figure 5 shows that the logistic GAM model again outperformed others, achieving an average accuracy of 67.4% and an F1-score of 64.6%. Elastic net and gradient boosting performed poorly, with average accuracies of 52.6% and 52.7%, respectively. In contrast, a simple majority classifier achieves an accuracy of 54% and F1-score of 35%. Graph-based features showed the best performance for Quests 2 and 3, with the logistic GAM achieving 70.7% and 73.8% accuracy, respectively. Logistic models offered more stable performance across all quests compared to gradient boosting and elastic net, likely due to their simplicity and suitability for the small dataset (61 samples). The logistic GAM model's ability to capture non-linear relationships without the risk of overfitting makes it preferable over more complex models like gradient boosting. Overall, while static features generally performed better, the superior results of graph-based features in Quests 2 and 3 suggest their potential benefits in predictive modeling.

5.2 Feature Importance Analysis

To explore the relationship between the extracted behavioral patterns and student learning outcomes, we focused on significant feature contributions from the highest-performing model, logistic GAM. We assessed the magnitude and direction of feature influences by examining matrix and spline coefficients. The magnitude of the coefficient matrix reflects the impact of feature contributions to the principal components. We assess the direction of influence for each component by examining the spline coefficients in the logistic GAM model. By comparing the ratio of positive to negative spline coefficients, we determine if a component has a positive or negative relationship with the outcome. An even ratio indicates a non-linear relationship.

Static and graph-based representations of collaborative problem-solving behaviors revealed nuanced, context-dependent relationships with student learning outcomes. In the tutorial phase, static features such as PC1—highlighting exploratory (EU), representation and formulation (RF), and social behaviors—were strongly predictive of learning gains, with higher coefficients compared to other activities. However, PC2, influenced by centrality and HITS authority measures, negatively correlated with outcomes, suggesting that frequent social interactions without sufficient reflection or problem solving can hinder progress.

As the game progressed, the relevance of static features diminished. For example, in Quest 2, PC1 captured problem representation (RF) and social behaviors, but their overall impact weakened. Conversely, graph-based metrics like graph diameter and centrality became more relevant in later stages, such as Quest 3, where PC1 positively correlated with learning gains through diverse behavioral transitions and central social interactions. PC2, however, showed a negative relationship, indicating that excessive reliance on "RF & EU" patterns—frequent answer formulation and exploration—may lead to ineffective engagement.

These findings suggest that the effectiveness of CPS behaviors evolves throughout gameplay. Early stages benefit from balanced exploration and collaboration, while later stages demand strategic transitions between behaviors. The combination of static and graph-based analyses provides a more comprehensive understanding of how CPS behaviors influence learning, emphasizing the importance of tailoring interventions to different stages of the learning process.

5.3 Viability of CSPM for CPS Behavior Analysis (RQ1)

A core motivation for devising a mapping between raw game trace data to CPS practices is to align observable in-game behaviors with well-established pedagogical constructs and understand students' CPS processes. We aimed to integrate theoretical insights from a CPS ontology with the granular, behaviorally rich data generated by students during collaborative game-based learning. By combining cluster analysis with Seq2Pat, we were able to extract various behavioral motifs for subsequent analysis. Findings revealed patterns that have varying degrees of relationship with student learning outcomes. Our mapping from these trace events to CPS practices increases the interpretability of game-based behaviors and can offer more insights into cognitive aspects of student behavior than can be gathered from raw trace logs alone. Our approach moves towards a scalable solution where raw trace logs can be transformed into meaningful educational constructs without manual intervention past the initial mapping stage.

Our mapping approach shows that in-game behaviors align well with cognitive CPS behaviors but struggle with social behaviors due to the abstraction introduced by clustering, which obscures discourse details. This suggests that the current system better supports cognitive aspects of CPS, and integrating discourse data in future analyses could enhance the assessment of social behaviors. Combining both data types could provide a more comprehensive view of students' CPS practices. The findings confirm that CSPM techniques like Seq2Pat, combined with clustering, effectively extract cognitive CPS behaviors. In-game actions were successfully mapped to four CPS ontology codes, highlighting the pedagogical value of structured pattern identification. This scalable approach automates the extraction of relevant data from trace logs, reducing manual intervention and allowing patterns linked to positive learning outcomes to guide future analyses. These methods could improve learning analytics by providing real-time feedback on cognitive engagement and helping educators tailor interventions to enhance problem-solving strategies.

5.4 Effectiveness of Extracted Patterns as Predictors (RQ2)

To evaluate the predictive utility of the extracted CPS patterns, we tested four machine-learning models, prioritizing explainability to better understand how specific student behaviors relate to learning outcomes. Among these models, the gradient boosting classifier underperformed compared to a simple majority classifier, indicating potential for future improvements. However, the logistic GAM emerged as a strong performer, demonstrating that the identified CPS patterns are viable predictors of student learning gains. Its predictive strength also highlighted the nuanced relationships between behavioral patterns and learning outcomes.

The logistic GAM achieved high accuracy and F1 scores across activities, performing best in Quest 1 with static patterns. This suggests that CPS patterns can effectively distinguish between higher- and lower-performing students. Static features proved especially valuable in earlier stages, capturing clear behavioral patterns associated with success. In later stages, graph-based features—capturing the sequence and transitions between behaviors—offered stronger predictive power by reflecting the dynamic nature of collaborative learning. This dual success underscores the complementary value of static and temporal data representations depending on the stage of gameplay.

Although individual behavioral patterns showed modest correlations with learning outcomes (less than 0.4), their aggregate predictive power was substantial. This suggests that while single patterns alone may be insufficient predictors, their combined relationships capture the complexity of student behavior. For example, in early tasks like Quest 1, static patterns effectively identified high-performing students, likely due to the structured nature of early activities. In contrast, later tasks (Quests 2 and 3) benefited from graph-based features that highlighted students' ability to strategically transition between CPS behaviors—an indicator of adaptive problem-solving strategies.

The logistic GAM's robust performance, even with a small dataset (61 samples), and its capacity to handle non-linear relationships make it particularly well-suited for educational contexts. Its interpretability further enhances its practical value, enabling targeted interventions to foster more effective collaborative problem-solving skills. These findings highlight the importance of adapting analysis techniques to different phases of learning, ensuring that interventions are aligned with evolving student behaviors.

5.5 Relationship Between Behavioral Patterns and Learning Outcomes (RQ3)

We investigated the relationship between static and graph-based representations of student CPS behaviors to distinguish higher- and lower-performing students. Both representations reveal a non-linear relationship between CPS behaviors and learning gains, suggesting interventions must be tailored to observed behaviors for maximum effectiveness. There may be behavioral frequency thresholds, saturation points, or counterproductive behaviors. Thresholds occur when certain behaviors need to happen frequently enough for significant learning gains, while saturation points arise when excessive behavior leads to plateaued or reduced gains due to over-reliance on a strategy.

In Quest 1, static patterns yielded the highest predictive performance across models, and its significant components showed a linear relationship with learning outcomes. The "EU & RF & EU" pattern positively correlated with learning, indicating that students engaging in exploration and answer formulation are more likely to perform well. Linear relationships, common in structured environments, suggest that Quest 1 is more predictable than other activities. Overall, static features reveal complex, synergistic relationships between CPS behaviors and learning outcomes. Tables 2 and 3 display the relationships between static and graph-based features, respectively, and learning gains.

Relationship of Static Variables with Learning Gain				
Name	**Component**	**Effect**	**CPS ontology**	**Coefficients**
Tutorial	PC1	Non-Linear	EU & social	1.579
			EU & social & EU	1.412
			social & EU & social & RF	1.337
Quest 1	PC1	Positive	EU & RF & EU	1.599
			EU & RF & EU & RF & EU	1.528
			EU & social & EU	1.465
	PC2	Negative	EU & RF	1.284
			EU & RF & EU & RF	1.223
			EU & social	1.040
Quest 2	PC1	Non-Linear	EU & social & EU	1.431
			EU & RF & EU	1.204
			social & RF & EU	1.147
	PC2	Non-Linear	EU & social & EU & RF & social	1.385
			RF & social	1.348
			EU & RF & EU	1.035
Quest 3	PC1	Non-Linear	RF & EU	1.320
			social & M & social	1.222
			EU	1.208
	PC2	Non-Linear	social	0.935
			RF & EU & social	0.931
			EU & RF & M & EU & M & EU	0.892

Table 2: Relationship between static variables and student learning gain.

Examining graph-based features (Table 3) reveals how CPS patterns' temporal interactions relate to student learning outcomes. For the Tutorial and Quest 3, graph-based features show linear relationships with performance, while Quests 1 and 2 exhibit non-linear relationships. The Tutorial's negative linear relationship indicates that frequent engagement in social and representation-formulation behaviors is linked to lower learning gains. Specifically, PC2 in the Tutorial involves actions like closing activities and submitting answers, with centrality and HITS authority measures suggesting students may rush through this section, leading to lower gains.

Quest 3 shows a positive linear relationship between learning outcomes and features such as graph diameter, average shortest path length, and centrality of social CPS patterns. This suggests that students displaying a diverse range of behavioral patterns and transitioning between them via key social behaviors are more likely to achieve higher performance. Conversely, frequent transitions into "RF & EU" patterns are negatively associated with learning gains, highlighting that too frequent RF & EU behaviors might imply that students quickly change their answers without in-depth contemplation or discussions. Social behaviors may be more beneficial than exploratory ones during this quest because we cannot ensure that social behaviors (i.e., conversations) are always productive and beneficial.

Quests 1 and 2 display significant non-linear relationships with learning outcomes. Principal components related to CPS behavior centrality imply that these behaviors serve as key transition points. The non-linear relationships suggest that the degree centrality of behaviors influences learning gains, with low centrality showing minimal impact and high centrality potentially indicating either effective or counterproductive engagement. Interventions should balance behavioral transitions to optimize learning outcomes.

Non-linear relationships between CPS patterns and student learning outcomes indicate that CPS behaviors' influence is context-dependent and complex. To understand their pedagogical impact, it is important to consider the presence, frequency, and timing of CPS behaviors. Static features predict early activities well but lose influence in later stages. In contrast, graph-based features, which capture the temporal dynamics of CPS behaviors, emphasize the importance of understanding not just the behaviors but also their timing and sequence. These findings highlight the need for both static and graph-based representations to accurately assess and predict student learning outcomes, advocating for a dynamic approach to student assessment that adapts to evolving interactions in game-based learning environments.

6 Limitations

Our work utilizes data from only 61 students from a collaborative game-based learning environment making it difficult to generalize the conclusions to other learning contexts and populations. Additionally, extracted patterns are specific to the clustering results for EcoJourneys. Application of these methods to new learning environments will require additional cluster and event topic analysis to generate meaningful descriptors of subsequent trace logs and would include re-training the initial embedding model on environment-specific trace log data. This in turn may also influence the types of patterns extracted and therefore would require an additional mapping between extracted patterns and relevant CPS practices. However, we believe that this necessitates a human-in-the-loop framework that allows for improved explainability and transparency for subsequent analysis. Finally, we employ fixed constraints in our CSPM algorithm which may lead to missing patterns that could be important at other time scales or frequencies. Future work should explore including different constraint types.

Relationship of Graph-Based Variables and Learning Gain					
Name	**Component**	**Effects**	**CPS ontology**	**Loadings**	**Features**
Tutorial	PC2	Negative	social	0.605	centrality
			RF	-0.393	centrality
			RF	-0.280	HITS authority
Quest 1	PC3	Non-Linear	EU & RF & EU	0.524	centrality
			EU & social & EU & social	0.432	centrality
			EU & RF & EU & RF & social	0.231	centrality
	PC5	Non-Linear		0.189	density
			social & EU & social & RF	0.187	centrality
			social & EU &RF & social & RF	0.146	centrality
	PC6	Non-Linear		0.275	density
			social & EU & social	0.267	centrality
			social & EU & social	0.203	HITS authority
Quest 2	PC1	Non-Linear		0.910	diameter
				0.398	avg. shortest path length
			social & EU & social	0.047	centrality
	PC2	Non-Linear	RF & social	0.464	HITS authority
			RF & social	0.406	centrality
			EU & social & EU & social	0.359	HITS hub
Quest 3	PC1	Positive		0.921	diameter
				0.381	avg. shortest path length
			social	0.025	centrality
	PC2	Negative	RF & EU	0.108	centrality
			RF & EU	0.090	centrality
			social	0.085	centrality

Table 3: Relationship between graph-based variables and student learning gain.

7 Conclusion & Future Work

Constraint-based sequential pattern mining can effectively extract collaborative problem-solving behaviors from raw game trace data. By mapping these behaviors to an existing CPS framework, we show that both static and temporal representations of student interactions can effectively predict student learning outcomes. The findings reveal that static features capture early learning behaviors well, while graph-based features, representing temporal dynamics, are more informative in later stages. The use of explainable models, such as logistic GAM, not only enhances prediction accuracy but also aids in understanding the non-linear and context-dependent relationships between CPS behaviors and learning outcomes. These insights offer potential for real-time educational interventions, fostering targeted strategies to improve collaborative problem-solving and overall student learning.

Future work should explore early prediction methods that look at the cumulative impact of each activity as students progress through the game. With this further analysis, researchers can achieve a more holistic understanding of the relationship between cumulative student behaviors across activities and their learning outcomes. We find that the trace log data can capture the cognitive dimension of students' use of CPS practices but fails to capture the social aspect of their interactions. Future work should identify and investigate discursive elements of students' collaborative interactions in conjunction with their in-game behaviors to provide a more complete understanding of CPS behaviors. Understanding the complex relationship between the cognitive and social dimensions of student behaviors can provide a means for more targeted intervention or personalization strategies for adaptive learning systems depending on whether students' behaviors are associated with high or low performance. This also opens pathways for designing supports that target students' specific needs to enhance their CPS processes. Finally, testing our approach with larger datasets and cross-context validation will help to assess the robustness and transferability of the framework. This will help to identify how the pattern extraction and analysis methods generalize across different learning contexts.

Acknowledgments

This research was supported by funding from the National Science Foundation (NSF) under Grants DRL-2112635, DRL-1561655, and DRL-1561486. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] Mehdi Acheli, Daniela Grigori, and Matthias Weidlich. 2021. Discovering and Analyzing Contextual Behavioral Patterns from Event Logs. *IEEE Transactions on Knowledge and Data Engineering* 34, 12 (2021), 5708–5721.

[2] Jessica Andrews-Todd and Carol M Forsyth. 2020. Exploring Social and Cognitive Dimensions of Collaborative Problem Solving in an Open Online Simulation-Based Task. *Computers in Human Behavior* 104 (2020), 105759.

[3] Jay Ayres, Jason Flannick, Johannes Gehrke, and Tomi Yiu. 2002. Sequential Pattern Mining Using a Bitmap Representation. In *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, 429–435.

[4] Roger Bakeman and Vicenç Quera. 2011. *Sequential Analysis and Observational Methods for the Behavioral Sciences.* Cambridge University Press, Cambridge, UK.

[5] Ryan S Baker, Taylor Martin, and Lisa M Rossi. 2016. Educational Data Mining and Learning Analytics. (2016), 379–396. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118956588.ch16

[6] Maria Bannert, Peter Reimann, and Christoph Sonnenberg. 2014. Process Mining Techniques for Analysing Patterns and Strategies in Students' Self-Regulated

Learning. *Metacognition and Learning* 9 (2014), 161–185.

[7] Arthur C Graesser, Peter W Foltz, Yigal Rosen, David Williamson Shaffer, Carol Forsyth, and Mae-Lynn Germany. 2018. Challenges of Assessing Collaborative Problem Solving. *Assessment and Teaching of 21st Century Skills: Research and Applications* (2018), 75–91.

[8] Daniel Carbonero, Jad Noueihed, Mark A Kramer, and John A White. 2023. Non-Negative Matrix Factorization for Analyzing State Dependent Neuronal Network Dynamics in Calcium Recordings. *BioRxiv* (2023).

[9] Huanyou Chai, Tianhui Hu, and Li Wu. 2023. Computer-Based Assessment of Collaborative Problem Solving Skills: A Systematic Review of Empirical Research. *Educational Research Review* (2023), 100591.

[10] Huili Dou, Jianhua Chen, and Junyi Zhou. 2023. Research on the Application of Data Mining in the Analysis of Students' Behavioral Intelligence in the Information Age. In *Proceedings of the 2023 International Conference on Information Education and Artificial Intelligence*. 859–864.

[11] Vegenshanti Dsilva, Johannes Schleiss, and Sebastian Stober. 2023. Trustworthy Academic Risk Prediction with Explainable Boosting Machines. In *Proceedings of the International Conference on Artificial Intelligence in Education*. Springer, 463–475.

[12] Ouafae El Aissaoui, Yasser El Alami El Madani, Lahcen Oughdir, Ahmed Dakkak, and Youssouf El Allioui. 2020. Mining Learners' Behaviors: An Approach Based on Educational Data Mining Techniques. In *Embedded Systems and Artificial Intelligence: Proceedings of ESAI 2019, Fez, Morocco*. Springer, 655–670.

[13] Andrew Emerson, Wookhee Min, Roger Azevedo, and James Lester. 2023. Early Prediction of Student Knowledge in Game-Based Learning with Distributed Representations of Assessment Questions. *British Journal of Educational Technology* 54, 1 (2023), 40–57.

[14] Sohom Ghosh, Shefali Yadav, Xin Wang, Bibhash Chakrabarty, and Serdar Kadıoğlu. 2022. Dichotomic Pattern Mining Integrated with Constraint Reasoning for Digital Behavior Analysis. *Frontiers in Artificial Intelligence* 5 (2022), 868085.

[15] Manuel J Gomez, José A Ruipérez-Valiente, Pedro A Martínez, and Yoon Jeon Kim. 2021. Applying Learning Analytics to Detect Sequences of Actions and Common Errors in a Geometry Game. *Sensors* 21, 4 (2021), 1025.

[16] Jiangang Hao, Lawrence Smith, Robert Mislevy, Alina von Davier, and Malcolm Bauer. 2016. Taming Log Files from Game/Simulation-Based Assessments: Data Models and Data Analysis Tools. *ETS Research Report Series* 2016, 1 (2016), 1–17.

[17] Xiuling He, Chenyang Wang, Yangyang Li, Zhipin Peng, and Jing Fang. 2022. Analysis of Group Online Collaborative Learning Based on Log Data and ICAP. In *Proceedings of the 2022 International Conference on Intelligent Education and Intelligent Research (IEIR)*. IEEE, 209–214.

[18] Friedrich Hesse, Esther Care, Juergen Buder, Kai Sassenberg, and Patrick Griffin. 2015. A Framework for Teachable Collaborative Problem Solving Skills. *Assessment and Teaching of 21st Century Skills: Methods and Approach* (2015), 37–56.

[19] Amin Hosseininasab, Willem-Jan van Hoeve, and Andre A Cire. 2019. Constraint-Based Sequential Pattern Mining with Decision Diagrams. In *Proceedings of the AAAI conference on Artificial Intelligence*, Vol. 33. 1495–1502.

[20] Deirdre Kerr, Jessica J Andrews, and Robert J Mislevy. 2016. The In-Task Assessment Framework for Behavioral Data. *The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (2016), 472–507.

[21] Lei Liu, Jiangang Hao, Alina A von Davier, Patrick Kyllonen, and Juan-Diego Zapata-Rivera. 2016. A Tough Nut to Crack: Measuring Collaborative Problem Solving. In *Handbook of Research on Technology Tools for Real-World Skill Development*. IGI Global, 344–359.

[22] Tongxi Liu and Maya Israel. 2022. Uncovering Students' Problem-Solving Processes in Game-Based Learning Environments. *Computers & Education* 182 (2022), 104462.

[23] Wenyi Lu, Joe Griffin, Troy D Sadler, James Laffey, and Sean P Goggins. 2023. Serious Game Analytics by Design: Feature Generation and Selection Using Game Telemetry and Game Metrics–Toward Predictive Model Construction. *Journal of Learning Analytics* 10, 1 (2023), 168–188.

[24] Abhishek Maheshwari, Amit Malhotra, Bhupendra Singh Hada, Monica Ranka, and Md Shaik Amzad Basha. 2024. Comparative Analysis of Machine Learning Models in Predicting Academic Outcomes: Insights and Implications for Educational Data Analytics. In *Proceedings of the 2024 International Conference on Smart Systems for Applications in Electrical Sciences (ICSSES)*. 1–7. https://doi.org/10.1109/ICSSES62373.2024.10561260

[25] Uwe Maier and Christian Klotz. 2022. Personalized Feedback in Digital Learning Environments: Classification Framework and Literature Review. *Computers and Education: Artificial Intelligence* 3 (2022), 100080.

[26] Sai Siddartha Maram, Johannes Pfau, Jennifer Villareale, Zhaoqing Teng, Jichen Zhu, and Magy Seif El-Nasr. 2023. Mining Player Behavior Patterns from Domain-Based Spatial Abstraction in Games. In *Proceedings of the 2023 IEEE Conference on Games (CoG)*. IEEE, 1–8.

[27] Fernando Marmolejo-Ramos, Mauricio Tejo, Marek Brabec, Jakub Kuzilek, Srecko Joksimovic, Vitomir Kovanovic, Jorge González, Thomas Kneib, Peter Bühlmann, Lucas Kook, et al. 2023. Distributional Regression Modeling via Generalized Additive Models for Location, Scale, and Shape: An Overview Through a Data Set from Learning Analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 13, 1 (2023), e1479.

[28] Richard E Mayer. 2019. Computer Games in Education. *Annual Review of Psychology* 70, 1 (2019), 531–549.

[29] Bradford W Mott, Robert G Taylor, Seung Y Lee, Jonathan P Rowe, Asmalina Saleh, Krista D Glazewski, Cindy E Hmelo-Silver, and James C Lester. 2019. Designing and Developing Interactive Narratives for Collaborative Problem-Based Learning. In *Interactive Storytelling: 12th International Conference on Interactive Digital Storytelling, ICIDS 2019, Little Cottonwood Canyon, UT, USA, November 19–22, 2019, Proceedings 12*. Springer, 86–100.

[30] OECD. 2017. *PISA 2015 Assessment and Analytical Framework*. OECD Publishing. 260 pages.

[31] Luc Paquette, Theodore Grant, Yingbin Zhang, Gautam Biswas, and Ryan Baker. 2021. Using Epistemic Networks to Analyze Self-Regulated Learning in an Open-ended Problem Solving Environment. In *Proceedings of the Advances in Quantitative Ethnography: Second International Conference, ICQE 2020, Malibu, CA, USA, February 1-3, 2021, Proceedings 2*. Springer, 185–201.

[32] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu. 2004. Mining Sequential Patterns by Pattern-Growth: The Prefixspan Approach. *IEEE Transactions on Knowledge and Data Engineering* 16, 11 (2004), 1424–1440.

[33] Elizabeth Rowe, Jodi Asbell-Clarke, Kathryn Cunningham, and Santiago Gasca. 2017. Assessing Implicit Computational Thinking in Zoombinis Gameplay: Pizza Pass, Fleens & Bubblewonder Abyss. In *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play*. 195–200.

[34] Asmalina Saleh, Cindy E Hmelo-Silver, Krista D Glazewski, Bradford Mott, Yuxin Chen, Jonathan P Rowe, and James C Lester. 2019. Collaborative Inquiry Play: A Design Case to Frame Integration of Collaborative Problem Solving with Story-Centric Games. *Information and Learning Sciences* 120, 9/10 (2019), 547–566.

[35] Yanjie Song, Jiaxin Cao, Yin Yang, and Chee-Kit Looi. 2022. Mapping Primary Students' Mobile Collaborative Inquiry-Based Learning Behaviours in Science Collaborative Problem Solving via Learning Analytics. *International Journal of Educational Research* 114 (2022), 101992.

[36] Ramakrishnan Srikant and Rakesh Agrawal. 1996. Mining Sequential Patterns: Generalizations and Performance Improvements. In *Proceedings of the International Conference on Extending Database Technology*. Springer, 1–17.

[37] Chen Sun, Valerie J Shute, Angela EB Stewart, Quinton Beck-White, Caroline R Reinhardt, Guojing Zhou, Nicholas Duran, and Sidney K D'Mello. 2022. The Relationship Between Collaborative Problem Solving Behaviors and Solution Outcomes in a Game-Based Learning Environment. *Computers in Human Behavior* 128 (2022), 107120.

[38] Mike Tissenbaum. 2020. I See What You Did There! Divergent Collaboration and Learner Transitions from Unproductive to Productive States in Open-Ended Inquiry. *Computers & Education* 145 (2020), 103739.

[39] Alina A von Davier, Jiangang Hao, Lei Liu, and Patrick Kyllonen. 2017. Interdisciplinary Research Agenda in Support of Assessment of Collaborative Problem Solving: Lessons Learned from Developing a Collaborative Science Assessment Prototype. *Computers in Human Behavior* 76 (2017), 631–640.

[40] Karen D Wang, Haoyu Liu, David DeLiema, Nick Haber, and Shima Salehi. 2024. Discovering Players' Problem-Solving Behavioral Characteristics in a Puzzle Game through Sequence Mining. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. 498–506.

[41] Xin Wang, Amin Hosseininasab, Pablo Colunga, Serdar Kadıoğlu, and Willem-Jan van Hoeve. 2022. Seq2Pat: Sequence-to-Pattern Generation for Constraint-Based Sequential Pattern Mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12665–12671.

[42] Günes Yavuz, Deringöl Yasemin, and Çigdem Arslan. 2017. Elementary School Students Perception Levels of Problem Solving Skills. *Universal Journal of Educational Research* 5, 11 (2017), 1896–1901.

[43] Dan Ye and Svoboda Pennisi. 2022. Using Trace Data to Enhance Students' Self-regulation: A Learning Analytics Perspective. *The Internet and Higher Education* 54 (2022), 100855.

[44] Mohammed J Zaki. 2000. Sequence Mining in Categorical Domains: Incorporating Constraints. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*. 422–429.

[45] Yingbin Zhang and Luc Paquette. 2023. Sequential Pattern Mining in Educational Data: The Application Context, Potential, Strengths, and Limitations. In *Educational Data Science: Essentials, Approaches, and Tendencies: Proactive Education Based on Empirical Big Data Evidence*. Springer, 219–254.