Exploring Policies for Dynamically Teaming Up Students through Log Data Simulation

Kexin Bella Yang
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
kexiny@cs.cmu.edu

Xuejian Wang
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
xuejianw@andrew.cmu.edu

Vanessa Echeverria
Escuela Superior Politécnica
del Litoral, ESPOL
Guayaquil, Ecuador
vanechev@espol.edu.ec

LuEttaMae Lawrence
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
llawrenc@andrew.cmu.edu

Kenneth Holstein
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
kjholste@cs.cmu.edu

Nikol Rummel
Ruhr-Universität Bochum
Universitätsstraße 150
D - 44801 Bochum
nikol.rummel@rub.de

Vincent Aleven
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
aleven@cs.cmu.edu

ABSTRACT

Constructing effective and well-balanced learning groups is important for collaborative learning. Past research explored how group formation policies affect learners' behaviors and performance. With the different classroom contexts, many group formation policies work in theory, yet their feasibility is rarely investigated in authentic class sessions. In the current work, we define *feasibility* as the ratio of students being able to find available partners that satisfy a given group formation policy. Informed by user-centered research in K-12 classrooms, we simulated pairing policies on historical data from an intelligent tutoring system (ITS), a process we refer to as *SimPairing*. As part of the process for designing a pairing orchestration tool, this study contributes insights into the feasibility of four dynamic pairing policies, and how the feasibility varies depending on parameters in the pairing policies or different classes. We found that on average, dynamically pairing students based on their in-the-moment wheel-spinning status can pair most struggling students, even with moderate constraints of restricted pairings. In addition, we found there is a trade-off between the required knowledge heterogeneity and policy feasibility. Furthermore, the feasibility of pairing policies can vary across different classes, suggesting a need for customization regarding pairing policies.

Keywords

Peer tutoring, Learning Group formation (LGF), Pairing Policies, CSCL

1. INTRODUCTION

Constructing effective, well-balanced learning groups is an important task in computer-supported collaborative learning (CSCL) [1–3]. The importance of learning group formation (LGF) has been validated empirically [4,5]. For instance, Webb et al.'s experiment proved that group composition had a major impact on the quality of group discussion and students' test scores, both during group work and subsequent individual tests [5]. The majority of existing approaches to LGF, do not support *dynamic* group formation [1]. Dynamic group formation refers to the process of groups "created on demand while various domain-specific restrictions have to be considered" [6], or can "adapt to and benefit from previous information about group members and their abilities" [7,8]. Compared to static, pre-planned LGF, the dynamic composition of groups allows for quick regrouping of learners based on up-to-date information regarding their progress and struggle. Dynamic group formation is an interesting issue, as researchers start envisioning more sophisticated and personalized classroom interactions [9] and more fluid social transitions (i.e., student social transitions that occur not all at the same time for everyone in the class) [10], that are more challenging to orchestrate.

In the context of an Intelligent Tutoring System (ITS) that supports both individual and collaborative learning, it is useful to investigate whether dynamically switching students between the two modes, as the need arises, can be effective and feasible. Pairing policies that work well in practice ideally have characteristics of both effectiveness and feasibility. By effective we mean that the pairing policy leads to students' reaching desired learning goals, and by feasible we mean that enough partners can be found under the given grouping policies (i.e., good policy coverage). Specifically, we defined *feasibility* as the percentage of students who can be teamed up under a given pairing policy.

The feasibility of LGF is an important issue to investigate in designing orchestration tools for teachers, and can be a central concern at the initial stage of tool design. This is because during the initial design stages we often do not yet have data to rigorously evaluate the effectiveness of LGF, given testing the LGF requires human resources of learners, instructors, materials resources of devices, systems, and a long time period. Additionally, an effective pairing policy that only covers a small percentage of students in a classroom may have limited influence for the whole class. Thus, the feasibility of LGF can be important in providing context for the potential coverage of LGF in a class.

Literature on LGF in collaborative learning is vast. Researchers have paired students based on gender [11,12], learning style [13–16], students' social network [17], and their intelligence or task proficiency [18–20]. Heterogeneous and homogeneous group formation are two main approaches in team formation, and many studies have demonstrated their effectiveness in CSCL [1,8,18,20–22]. Students' knowledge level is argued to be the most suitable and important attribute to form educational groups [8]. Prior work has also used machine learning or other algorithms

to incorporate multiple factors for optimizing team formation [23–26]. However, the literature on LGF provides little insight into the feasibility of these LGF policies, especially in the ITS context.

Evaluating the feasibility of LGF policies offline, prior to implementing them in real classrooms, is challenging, given a lack of readily accessible approaches. To address this problem, we adopt a process we call "*SimPairing*", to simulate pairing policies on authentic data and evaluate their feasibility. In this process, we used transaction data from several classes of students using an ITS, collected from classroom studies conducted in U.S. middle schools. We computed and analyzed how the feasibility of several LGF policies (described below) changed as each class progressed, and how the feasibility varied across different classes. Replaying historical data to simulate possible futures (e.g., Replay Enactment [27]), has been used as a method by researchers to design tools with similar data-driven, human-centered approaches [28]. Diana et al. [29], for instance, used machine learning (ridge regression) to predict students' grades based on historical data in CS education (i.e., programming). Based on these *predicted grades* and simulated *students' "helped" status,* they determined which students needed help and which may be able to provide help. They then used a network graph of code-state to search for potential peer tutors who shared a common ancestor node with the tutee. They found that grouping low-performing students together and using better model features can increase the number of students helped. Their findings suggest that using low-level log data to group and match low-performing students with a peer tutor may be an effective way to increase the amount of help given in a classroom. In contrast, we simulated different policies selected based on literature and teachers' common practice revealed in user-centered research with K-12 teachers [10,30,31], in a mathematics education context.

The current work is, to the best of our knowledge, *the first to look at dynamic pairing policies that consider students' in-the-moment wheel-spinning status.* Identifying students who are unproductively struggling, yet failing to master the skill, (i.e., wheel spinning) is a first step to getting them unstuck [32]. While there has been significant work on modeling and predicting wheel spinning [33–35], little work has been dedicated to developing interventions to get them unstuck, with a few recent exceptions [36,37]. While a typical classroom has students who are struggling on problems and those who have excelled on the same problem, the latter students' expertise is rarely utilized. Instead, often the only source of help is the instructor, who is likely unable to help all the students who need help within the time constraints of the class period [38]. Peer tutoring (i.e., pairing a struggling student with a peer tutor) could be an effective way to help get struggling students unstuck when the instructor has their hands full.

Lastly, instead of prescribing *a specific grouping criterion, our work envisions that instructors will customize pairing policies and parameters to their classroom contexts, which prior work argued to be especially helpful in the LGF process [1,8,30,39].* Amara et al. found that most of the proposed LGF solutions do not allow instructors to customize the grouping process [1]. They argued that it is less helpful to apply a grouping solution for all types of learners, and more useful to leave the choice to instructors. Instructors can then form groups according to different learning objectives, learners' needs, activity types, and customize the LGF process according to location and time [1]. Similarly, Echeverria et al. envision adaptability in an orchestration system, which "enables teachers to select the best pairing policies based on their particular goals, needs, and classroom dynamics" [30], to be helpful for different classrooms. In the current investigation, three of the four policies we studied involve an adjustable pairing threshold or parameter, which we simulated with various values.

In sum, the current work investigates the feasibility of four dynamic LGF policies derived from user research with math teachers. We investigate from three angles: overall session simulation, class-level variance, and session-level contrasting cases. This work contributes to the feasibility results of the dynamic pairing policies, recommendations for orchestration tool design, and highlights future work regarding tools supporting dynamic LGF.

2. STUDY CONTEXT

2.1 Intelligent Tutoring Systems

This study used student transaction data collected from classroom studies in U.S. middle schools (dataset link). This data logged students' interaction with an ITS called Lynnette, which offers guided practice to students in basic equation solving. ITS (also called AI-tutors) are increasingly common in K-12 classrooms to help teachers more effectively personalize instruction [40]. As shown in Figure 1, Lynnette provides step-by-step guidance, in the form of adaptive hints, correctness feedback, and error specific messages. Lynnette supports personalized mastery learning, and has been proven to improve students' equation-solving skills in several classroom studies [41–43].

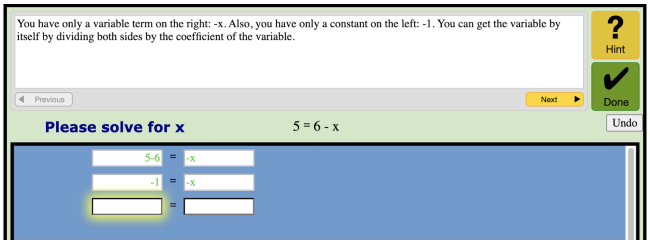

Figure 1. Example student interface for the ITS, Lynnette

The transaction data logs detailed events by the timestamp of students' interaction with the ITS, including but not limited to actions they take (e.g., requesting hints or attempting a step), knowledge components (KC) that a transaction involves, and skill mastery, calculated based on Bayesian Knowledge Tracing (BKT) student model, a two-state Hidden Markov Model. BKT is a popular student model that has been successful for various applications in the educational technology literature (e.g. [44]).

The current work lays a foundation to (in the future) use Lynnette in combination with a second ITS, APTA, which extends Lynnette's functionality to support reciprocal peer tutoring. APTA allows two students to respectively take the role of tutor and tutee. In APTA, the tutee can seek help from their partner, while the tutor can see the tutee's progress, and help them to make progress with the math problem at hand. APTA supports the peer tutor in tutoring the tutee. Classroom studies with APTA have demonstrated that its adaptive support can improve the quality of help peer tutors give and improve students' domain learning [45,46]. A future effort for the current work is to implement feasible student pairing policies in an orchestration tool, to support teachers in dynamically pairing students to work collaboratively in APTA. Such a tool plays a key role in our vision for the smart classroom of the future, in which students alternate fluidly between individual and collaborative learning.

2.2 Wheel-Spinning Detector

Detectors have been developed to detect student behaviors of interest (e.g., gaming the system, struggling) from the transaction data. Such detectors have been used to design dashboards or tools that can alert teachers of certain student status (e.g., [9]). In our policies 1 and 2, we paired students based on their struggle status indicated by a wheel-spinning detector.

Wheel-spinning, as defined by Beck and Gong, denotes students who are failing to master a specific skill after many attempts in an intelligent tutoring system [32]. We utilized a detector that adopted the same criterion as defined by Beck and Gong [32]. The detector is embedded in LearnSphere, (i.e., a large learning analytics infrastructure) [47]. The detector considers students who have over ten practice opportunities yet still failing to reach a skill mastery on a specific knowledge component (KC) of above 0.95, to be wheel-spinning on this KC [9]. Such prolonged repeated struggles are likely to be an inefficient use of time for students [32] and may contribute to a lack of motivation for future learning [36]. Wheel-spinning is one type of unproductive struggle, and we use *struggling* and *wheel-spinning* interchangeably in this paper.

3. METHODS

We evaluated how feasible four pairing policies (described below) are, based on simulation with historical transaction data from Lynnette. We applied each pairing policy to data from each class session. For every minute in a session, we calculated the percentage of students who met the policy's criterion for being teamed up. Based on this calculation, we evaluated policy feasibility using two measures, FI_1 and FI_2, defined in 3.2. In the simulation process, we did not make assumptions about how long simulated collaboration episodes would last. We foresee that in any of these episodes, students will be given the task of collaboratively solving several math problems; it is hard to predict how long that will take them. We thus did not simulate taking tutors or tutees out of the pool of students available for teaming up, or returning them to this pool, at the beginning and end of collaborative episodes, respectively. Although this simplification might introduce some inaccuracy into the simulation results, it may be hard to do better. As well, the asymmetric roles that paired-up students have in the pairing policies may limit the inaccuracy. For example, simultaneously keeping a struggling student and a non-struggling in the pool instead of taking them both out might have offsetting effects in terms of feasibility.

Our simulation involved four pairing policies, namely:

Policy 1 - Struggle with Non-Struggle: Pairing students who are wheel-spinning (unproductive struggle) with students who are not wheel-spinning.

Policy 2 - Pairing with Restriction: Pairing students who are wheel-spinning with those who are not wheel-spinning, with a varying pairing restriction (PR) rate β. The PR rate simulates restrictions regarding who can collaborate with whom, which in real life would be provided by the teacher.

Policy 3 - Knowledge Difference Pairing: Pairing students whose knowledge levels (as measured by the tutor's BKT) differ by *more* than a certain threshold α.

Policy 4 - Knowledge Similarity Pairing: Pairing students whose knowledge levels (as measured by the tutor's BKT) differ by *less* than a certain ceiling γ.

The distinction in these policies aligns with Amara et al.'s categorization for dynamic group formation [1]: *intra-session* and *inter-session* grouping. Intra-session grouping allows for changing group members during the learning process, which is useful, for example, for synchronous mobile collaborative learning [1]. In *inter-session grouping,* groups are formed only before starting or after ending the learning process. Specifically, policies 1 and 2 fall under *intra-session* grouping since we simulated pairing students based on their in-the-moment struggle. These two policies also concern *fluid social transitions [10]*, since the students in a given class may transition from individual to collaborative learning at different times. Our pairing policies 3 and 4 concern inter-session grouping, and pair students based on their initial knowledge level. To apply these policies, teachers or the tutoring system would assess students' knowledge level, prior to (or at the beginning of) a class session.

The research questions we aim to answer are:

RQ1: Based on a pairing simulation done with students' historical transaction data, how feasible are the four pairing policies?

RQ2: How does varying the parameters in the pairing policies affect the feasibility of pairing students?

RQ3: Does the feasibility of the pairing policies vary for different classes or sessions, if so, how?

3.1 The Four Pairing Policies

3.1.1 Policy 1: Struggle with Non-Struggle

Description. Policy 1 utilizes the struggle detector (section 2.2) to pair students who are wheel-spinning with those who are not. The struggle detector assumes students' wheel-spinning status to be a binary value for a given timestamp. Inspired by the work of Diana et al. [29], we categorized students in the *Struggle Pool* if they were wheel-spinning on at least one KC, indicating they could need help from a partner. Students not wheel-spinning on *any* KC were categorized in the *Tutor Pool* and considered as available tutors. We simulated pairing students in the *Struggle Pool* with students in the *Tutor Pool*. To determine the feasibility of this policy, we calculated the percentage of struggling students who had a potential partner (for more detail, see below).

Rationale. Literature suggests that when students are wheel-spinning, giving them more of the same type of math problems to solve may not be productive [36]. When wheel-spinning, students would likely benefit from instructor attention or extra instruction. However, prior user research in the classroom (e.g. [9]) found that teachers often cannot help all struggling students. In this case, wheel-spinning students may benefit from a peer tutor's help, which leads to a policy that seeks to dynamically find them partners [36].

3.1.2 Policy 2: Pairing with Restriction

Description. Policy 2 is an extension to Policy 1, where we pair a struggling student with a non-struggling student, while enforcing a constraint that not all students are eligible for teaming up. The proportion of ineligible students is captured as the Pairing Restriction (PR) rate. The PR rate is used to simulate situations where the teacher prefers that certain students do not work together. Specifically, we simulate pairing students in the *Struggle Pool* with students in the *Tutor Pool,* while enforcing the restriction that β% ($0 < \beta < 1$, step = 0.1) of students in the *Tutor Pool* are ineligible as partners. For example, a PR rate β of 0.2 means 20% of the students in *Tutor Pool* have been restricted

from working with any students in *Struggle Pool*. It is important to know how these restrictions affect the feasibility of the policies.

Rationale. We designed this policy based on results found in a survey we conducted with 54 middle-school math teachers on their pairing preferences in collaborative learning [31] and semi-structured interviews conducted with middle school teachers. Teachers expressed a desire to set constraints so that certain pairs of students are restricted from working together. Previous studies and user research by Olsen et al. and Echeverria et al. also informed the idea of ruling out certain pairings in advance [10,30]. Such restrictions usually arise from information or concerns teachers have about their students' traits, behaviors and interpersonal relationships [8].

3.1.3 Policy 3: Knowledge Difference Pairing

Description. In Policy 3, we pair students who have *different* Initial Knowledge (IK) levels. In a practical scenario, teachers may assess students' knowledge through quizzes or exams. Alternatively, if the classrooms use ITS, teachers may have students practice several math questions individually, prior to transitioning into collaborative learning activities.

To simulate this policy without having pre-assessment data, we used data from the tutoring sessions (captured in the log data) to compute students' IK levels. Specifically, we computed a student's IK for each KC, as the average mastery for their first *three* opportunities for this KC. The reason is we want to use up only a small portion of the data from the tutoring session, so the measure represents initial knowledge. In our datasets, three opportunities generally fall in the first quartile (25%) of students' total number of opportunities for any given KC. Another reason we chose the cutoff of three is a previous EDM study with ASSISTments data showed student learning often appeared to occur, after students have had *ten* opportunities with the target knowledge [48]. Thus one may assume learners to have little learning on their first three times in transaction data practicing a KC. A student's overall IK S_j $(j \in N)$ is calculated as the average of their IK across KCs. To more accurately calculate students' *IK*, we limit our simulation to sessions that practiced the first (i.e., the most basic) level of KCs, involving 25 sessions.

KD was the difference between two students' IK, and denoted as S_{jk} $(j,k \in N)$, which was calculated as the absolute value of differences between two students' IK:

$$KD(S_{jk}) = |IK(S_j) - IK(S_k)|$$

Inspired by Huang and Wu's work that proposed a clustering LGF method that considers a threshold of learner heterogeneity [49], this work similarly considers a KD threshold. For this policy, the required KD of two students (S1, S2) should be a *minimum* of α ($0 < \alpha < 1$, step = 0.1) for them to be eligible to pair up.

Rationale. The heterogenous pairing policy was informed by findings from user research with math teachers. In the survey conducted with 54 math teachers, we found the most common way teachers paired students was pairing those who have a different level of knowledge (67%, N = 34) [31]. In our study, we use students' *mastery of knowledge components* (i.e., targeted math skills) calculated based on the BKT model to represent students' knowledge. In a systematic literature review on LGF in CSCL, Maqtary et al. found the knowledge level is the most commonly used attribute in LGF, which they claim to be the most suitable and important attribute to form educational groups because of its effects on the group process [8].

There is a range of research that shows heterogeneous grouping can promote positive interdependence, better group performance, and effective interactions [1,49–52]. Heterogeneous group composition not only enhances elaborative thinking, but also leads learners to deeper understanding, better reasoning abilities, and accuracy in long-term retention [49,50]. Research also suggests that collaborative learning with heterogeneous group composition by characteristics such as gender, ability, achievement, social-economic status (SES), or race, can be beneficial [51].

3.1.4 Policy 4: Knowledge Similarity Pairing

Description. Policy 4 is analogous to Policy 3, with the same definition of KD and IK as in Section 3.1.3. To pair students with *similar* knowledge, using the same calculation as Policy 3, this policy simulated pairing students that have *a small* KD. To be eligible for students to form a pair under this policy, the KD of two students (S1, S2) should be *less than or equal to* γ ($0 < \gamma < 1$, step = 0.1). For example, when γ = 0.2, two students with knowledge of 0.6 and 0.75 (KD = 0.15, below γ) would be eligible to pair, but another pair with knowledge of respectively 0.5 and 0.8 (KD = 0.3, above γ) would not be eligible.

Rationale. Policy 4 was inspired by prior literature and informed by user research. Literature suggests that homogenous groups can be beneficial for students' learning. For example, Fuchs et al. found homogenous dyads generated greater cognitive conflict and produced better quality work than heterogeneous groups [22]. Additionally, among 54 teachers we surveyed, 43% reported that they pair students with a similar level of knowledge [31]. This was the third most popular grouping method that teachers commonly adopt (43%, N = 23), following strategies of pairing students with different knowledge (Policy 3) and pairing students randomly *[31]*.

3.2 Metrics

In this section, we describe the metrics to evaluate the pairing policies. We discuss how prior work informed the metric definitions, and how different metrics could be suitable to evaluate different policies. We build on Diana et al.'s work [29], who defined an Efficiency Index (EI) as a measure of a pairing algorithm's performance, specifically:

$$EI = \frac{LowPerformingStudentsHelped/BeingHelped}{LowPerformingStudents}$$

We adapted EI into two metrics of interest for our pairing policies: Feasibility Index 1 and 2. FI_1 is the percentage of students who can be paired among all struggling students in a session.

$$Feasibility\ Index\ -\ 1\ (FI_1) = \frac{StrugglingStudentsCouldBeHelped}{TotalStrugglingStudents}$$

FI_2 is the ratio of paired students among all the students in a session.

$$Feasibility\ Index\ -\ 2\ (FI_2) = \frac{StudentsPaired}{TotalStudents}$$

For Policies 1 and 2: Given the goal to pair all struggling students in the session, FI_1 was a suitable measure for policy feasibility, showing what percentage of students who are wheel-spinning can get help. *For Policies 3 and 4:* Given the goal to pair all students in the session who satisfied a certain KD, FI_2 was a suitable measure for policy feasibility, as it calculated the percentage of the paired students out of the total students.

3.3 SimPairing Approach

There are three main steps in *SimPairing*: 1) data cleaning and preprocessing, 2) policy simulation, and 3) policy evaluation. The data cleaning and preprocessing step consists of clustering student transaction data into meaningful class sessions based on meta-data (e.g., student transaction timestamp, classes), and examining the distribution of students per class session to detect outliers. The policy simulation step takes the preprocessed transactional data and applies a pairing policy to class sessions. In the policy evaluation step, we computed the policy feasibility based on the simulation results, using the corresponding feasibility index (FI_1 or FI_2). We also observed how the FI changed by varying the parameters (i.e., KD, and PR rate).

4. ANALYSIS AND RESULTS

4.1 Data Cleaning and Preprocessing

We first clustered student transaction data into meaningful class sessions, based on timestamp, student ID, and class. We visualized student engagement for all class sessions based on transaction data, which allowed us to ensure that the sessions we analyzed had a continuous student interaction with the system, and helped us check for outliers (e.g., unusually short sessions). We excluded four outlier sessions: 2 sessions that had only 1 student, 2 sessions that lasted less than 15 min, as sessions commonly lasted 40 minutes or more.

Transaction data of a total of 68 sessions, from six middle school math classes, collected from 2013 to 2014 were used for policy simulation. It consists of 894 students and 197,234 rows of transactions. The average number of students in a session was 13 (*Min* = 5, *Max* = 24, *SD* = 25.3); the average duration of class session was 41.9 minutes (*Min* = 10, *Max* = 81, *SD* = 9.42); the average number of sessions in a class was 11 (*Min* = 3, *Max* = 23, *SD* = 9.33).

4.2 Overall SimPairing Analysis

In this section, we present, for each policy, the SimPairing analysis and the results. The goal for this analysis was to evaluate the overall feasibility of the four pairing policies (RQ1) and see how the feasibility depends on policy parameters (RQ2).

4.2.1 Policy 1: Struggle with Non-Struggle

We simulated Policy 1 for every minute in a given class session, which returned the number of struggling students who did or did not have a potential partner. Based on this we calculated the FI_1 for every minute in a class session. We then averaged FI_1 across the length of each class session, to obtain an average FI_1 for a given session. We refer to it as the Average Number of Struggling Students (ANSS). We then took the average of the ANSS across all sessions, to obtain an overall simulation result for all 68 sessions. Figure 2 (green area) shows the average FI_1 for all sessions was 0.94 (*SD* = 0.007). Thus, on average, across time, 94% of struggling students could be paired with a partner who was not struggling.

4.2.2 Policy 2: Pairing with Restriction

The Policy 2 simulation process is similar to Policy 1, with the addition of enforcing a varying PR rate. PR rate specifies a percentage of students in *Tutor Pool* as restricted from partnering with students in the *Struggle Pool*. We computed FI_1 with varying PR rates. As shown in Figure 2 (white area), FI_1 dropped as the PR rate increased, as expected. However, even with a relatively high PR rate of, for example, 0.4, meaning, 40% of non-struggling students are restricted from working with struggling students, we still get a high average FI_1 of around 0.80, (i.e., 80% of struggling students could be paired). The simulation result means that teachers can afford to set moderate restrictions for pairings, without compromising too much of the pairing policy's feasibility.

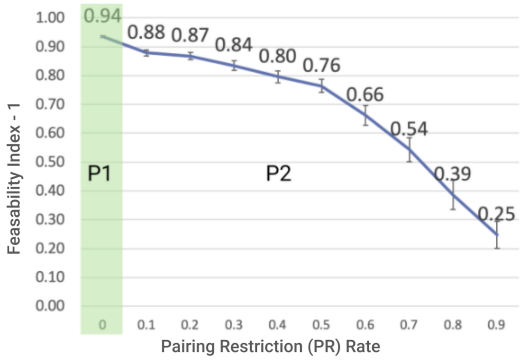

Figure 2. FI_1 for Policies 1 and 2

4.2.3 Policy 3: Knowledge Difference Pairing

Policy 3 requires students to be above a given *minimum* distance in their IK to be eligible for pairing up. We simulated this policy by computing FI_2 with varying values for the KD distance threshold α. We simulated these sessions to calculate the FI_2. As in Figure 3 (blue line), FI_2 dropped rather quickly as the required knowledge distance threshold went up. For example, the simulation results show that if we want to ensure an average of 80% of paired ratio, the KD threshold should be set to less than approximately 0.1 (i.e., a very strict bar).

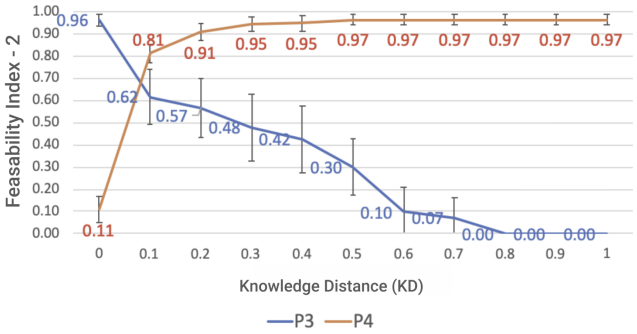

Figure 3. FI_2 for Policies 3 and 4

4.2.4 Policy 4: Knowledge Similarity Pairing

Policy 4 requires students to be below a *maximum* distance in their IK to be eligible for being paired up. We simulated this policy and computed FI_2 with different values for the KD distance ceiling γ. We found that this policy would work well even with a low, strict ceiling for the knowledge distance (Figure 3, red line). For example, when γ was 0.1, (i.e., two students' knowledge distance can be at most 0.1 for them to be teamed up), the average FI_2 was still 0.81 (*SD* = 0.08) across the class sessions involved. When γ was set to above 0.3, 95% of students in class could find an eligible partner.

4.3 Class-Level Variance Analysis

We explored how the four pairing policies worked for different classes and whether the pairing policies should be adapted to class-level differences (RQ3).

4.3.1 Class-level Differences

Based on Echeverria et al.'s insight that pairing support for teachers should ideally be adaptable to different classroom contexts [30], we analyzed, first, if there were systematic differences between different classroom contexts, and second, if these differences relate to policy feasibility differences. The main context variables taken into account by our pairing policies are *students' struggle status* (Policies 1 and 2) and *initial knowledge* (Policies 3 and 4). We thus analyzed if the classes had different *struggle statuses* and *initial knowledge (IK)*.

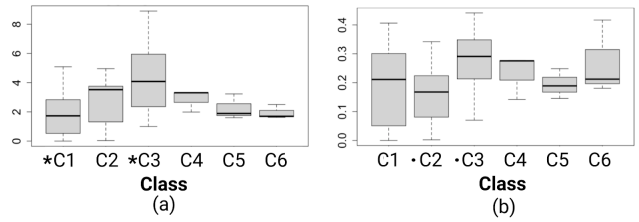

Figure 4. ANSS (a) and ASSRs (b) for Six Classes

***Student struggle status*:** We first calculated the number of students in wheel-spinning status for every minute within each class session. We then computed ANSS (section 4.2.1), and the *Average struggling students ratio (ASSR)= ANSS / total number of students in the session*. Histograms for ANSS and ASSR for all sessions show they follow the normal distribution. We conducted one-way ANOVAs, respectively taking the *ANSS* and *ASSR* as outcome variables and *Class* as the explanatory variable. The results showed a significant difference for ANSS among classes (Figure 4, a) [$F(5,62) = 4.34$, $p < 0.001$]. *Post hoc* Tukey tests showed C3 and C1 have significant differences (*diff* = 2.55, $p < 0.001$). All *post hoc* pairwise tests conducted in this study were corrected for multiple comparisons. The ANOVA result indicated that the classes differed with marginal significance [$F(5,62) = 1.94$, $p < 0.1$] (Figure 4, b). *Post hoc* Tukey tests showed a marginal difference in the ASRR between C3 and C2 (*diff* = 0.10, $p = 0.08$). Thus, there were class-level differences with respect to students' struggle status.

Initial Knowledge: We calculated each student's IK for all KCs involved (defined in section 3.1.3). The histogram for all students' IK shows it follows the normal distribution. We then conducted one-way ANOVAs using *IK* as the outcome variable, and *Class* as the categorical explanatory variable. The results indicated a significant effect of classes on IK for the six classes [$F(5, 320) = 5.895$, $p < 0.05$], and the IK for the six classes were not all equal. From the *post hoc* Tukey tests comparing knowledge level between each pair of the classes, we saw significant differences between classes C2 and C1 (*diff* = 0.12, $p < 0.05$), C3 and C2 (*diff* = -0.095, $p < 0.05$), and C4 and C2 (*diff* = -0.189, $p < 0.05$). C2 had the highest median of student IK (Figure 5), and a significantly higher level of IK than C1 and C3, and C4.

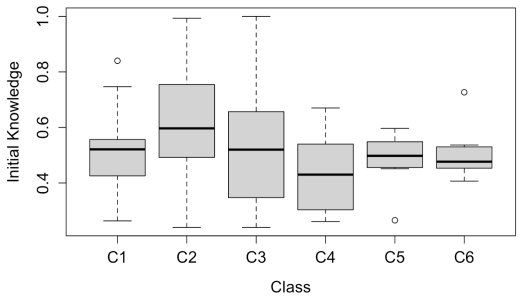

Figure 5. Initial Knowledge for Six Classes

Having characterized struggle and IK at a class level, we compare the policies' feasibility across classes.

4.3.2 Policies 1 and 2

Policy 1 had an average FI_1 above 0.85 (Figure 6, green area). We statistically compare if Policy 1 behaved differently for each class and see whether this policy should be adaptable for each class. Using session as the unit of analysis, we conducted a one-way ANOVA using the FI_1 for each session as the outcome variable, and *Class* as the categorical explanatory variable. The results indicated that there was not a significant effect of class on FI_1 [$F(5,62) = 1.24$, $p = 0.30$]. This result showed that Policy 1 was relatively consistent across the six classes, suggesting that Policy 1 may not need to be adaptable to classes.

For Policy 2, with increasing PR rate, the FI_1 decreases at a different speed for different classes, indicating some degree of class-level difference (Figure 6, white area). We conducted ANCOVAs with *Class* being the categorical explanatory variable, the *PR rate* as the quantitative explanatory variable, and FI_1 being the quantitative outcome variable. We first compared the model with and without a Class × PR rate interaction term. The model comparison result showed no evidence of an interaction effect among explanatory variables ($F = 1.63$, $p = 0.15$). We thus perform ANCOVA using an additive model. Results indicated there were eight pairs of classes that had significant differences in FI_1 for this pairing policy ($p < 0.05$). The eight pairs were C1-C2, C1-C3, C1-C6, C2-C3, C2-C6, C3-C5, C4-C6, and C5-C6.

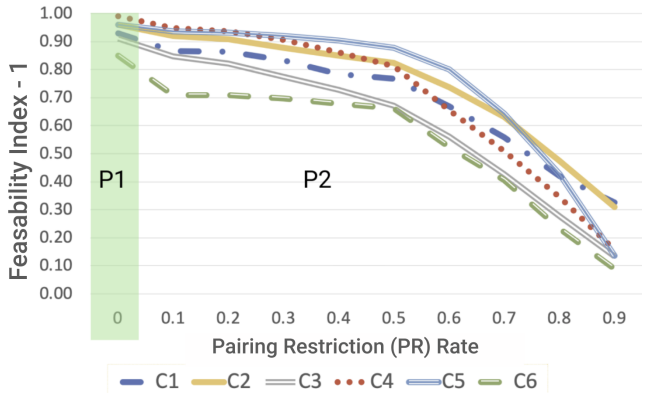

Figure 6. FI_1 of Policies 1 and 2 for Six Classes

Next, we looked at possible relations between the class-level feasibility variance of policies 1 and 2, and the class-level differences in *struggle status* (section 4.3.1). We found that for classes that differed with respect to the number of struggling students (C3-C2) and the average ratio of student struggle (C3-C1), the feasibility of Policy 2 tended to differ as well. This finding suggests that 1) Policy 2 may benefit from being adaptable to class-level characteristics, and 2) variables characterizing a class's struggle status (e.g., ANSS and ASSR) may have value in indicating how Policy 2 should be adaptable. On the other hand, the feasibility of Policy 2 was different in Class 6 compared to all other classes except C3, yet Class 6 did not differ in number or ratio of struggle students from other classes. Thus, students' struggle status alone may not provide enough information to fully decide whether and how P2 should be adaptable.

4.3.3 Policies 3 and 4

For Policy 3, the classes shared a downward trend in FI_2 with different slopes for each class (Figure 7). For example, when the

KD was 0.1, we saw the FI_2 values for class 6 (green dotted-line) drop to as low as 50%, but the other five classes have FI_2 above 75%. This shows that policy feasibility may be differently affected by the knowledge heterogeneity threshold in each class.

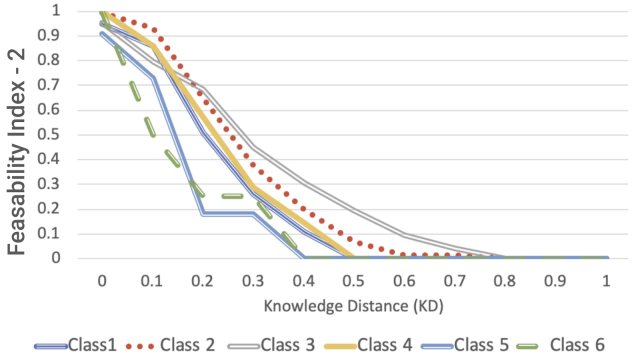

Figure 7. FI_2 of Pairing Policy 3 for Six Different Classes

To test whether Policy 3 behaved differently for each class, we conducted ANCOVAs with *Class* as a categorical explanatory variable, *KD* as a quantitative explanatory variable, and FI_2 in each session as the outcome variable. We first compared the model with and without a Class × KD interaction term. Results indicated no evidence supporting the interaction effect ($F = 0.81$, $p = 0.54$). We performed an ANCOVA using an additive model. Results indicated that three pairs of classes had significant differences in FI_2 ($p < 0.05$), and that two pairs of classes were marginally different ($p < 0.1$). They were C1-C3, C3-C5, C3-C6 ($p < 0.05$) and C2-C5, C2-C6 ($p < 0.1$) .

For Policy 4 (Figure 8), the six classes were more convergent and clustered closer together than Policy 3 (Figure 7). This indicated the class level difference may not be as strong as that in Policy 3, which our ANCOVA tests confirmed. Similar to Policy 3, we compared whether Policy 4 behaved differently for each class. We conducted an ANCOVA, with *Class* as a categorical explanatory variable, *KD* as a quantitative explanatory variable, and FI_2 as the outcome variable. We first compared the model with and without a Class × KD interaction term. No evidence supporting interaction effect among explanatory variables ($F = 0.13$, $p = 0.99$). We then performed ANCOVA using an additive model. Results indicated that there were no significant differences in FI_2 ($p > 0.05$) for Policy 4. We confirmed a smaller class-level difference as compared to Policy 3, in KD's effect on policy feasibility. From this result, we conclude Policy 4 performed quite consistently across classes, and no significant evidence showed that Policy 4 should be adaptable to classes.

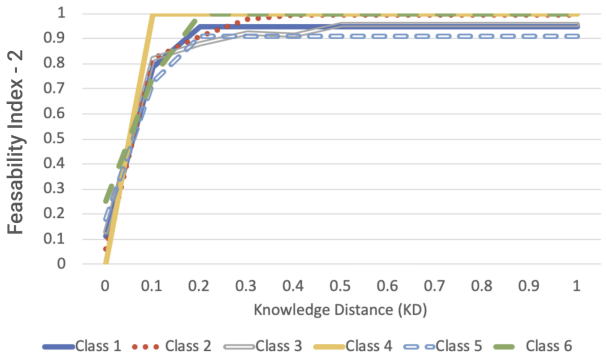

Figure 8. FI_2 of Policy 4 for Six Different Classes

Analogous to policies 1 and 2, we then looked at relations between feasibility variance for policies 3 and 4 and class-level IK characteristics in Section 4.3.1. We observed significant differences in IK between C2-C1, C3-C2, and C4-C2. However, the differences in IK for two classes cannot accurately predict whether they had different feasibility in Policy 3 and Policy 4, and other classroom characteristics may be needed to accurately represent the class-variance of feasibility.

4.4 Analysis of Contrasting Cases

We conducted a case study to understand how policies may perform dynamically (e.g., across every minute during class time) and differently in different class sessions (RQ3). For every policy, we selected a *typical* case and an *extreme* case in terms of the policy feasibility simulation results. For the *typical* case for all four policies, we selected a session (Session 1, C1) that had an average length of time (i.e., 41 minutes), an average number of students (i.e. 13 students). In the session, policies performed typically (as by visually comparing the simulation results of each policy for all sessions). As for the *extreme case*, we examined the simulation results for each policy on each session, and identified different sessions where each policy performed surprisingly or differently from the common trend. The extreme case can be a worst case scenario (Policies 1, 2 and 3) or a case that works surprisingly well (Policy 4). Below, we present the analysis and results for these contrasting cases for each policy.

4.4.1 Policy 1

In Policy 1, we chose the extreme case (Session 19, C3) as it was a session that this policy has the worst performance on, and thus it had the most different FI_1 trend, from examining visualizations of FI_1 for all sessions involved. We compare the typical case and extreme case by first contextualizing the struggle status of the two cases, and comparing the visualization of feasibility (for each minute) in the two sessions. Figure 9 depicts the *ratio of struggling students* (among all students in the class session) for the contrasting cases. For Policy 1 simulation (Figure 10), we obtained, for every minute in the class session, three values regarding policy feasibility: the number of students who were not wheel-spinning on any KCs (green bar), the number of students who were struggling, and *had* a potential partner (yellow bar), and the number of students who were struggling and *did not have* a potential partner (red bar).

Typical Case. In the typical case, students started to struggle after 10 minutes, as shown in Figure 9 (a) and Figure 10 (a). In all instances of wheel-spinning, a potential partner was available (i.e., $FI_1 = 1$ for any minute in this session). The typical case aligned with the overall simulation results from Policy 1, which showed that, for an average class, most struggling students could find potential partners, the minute they struggled.

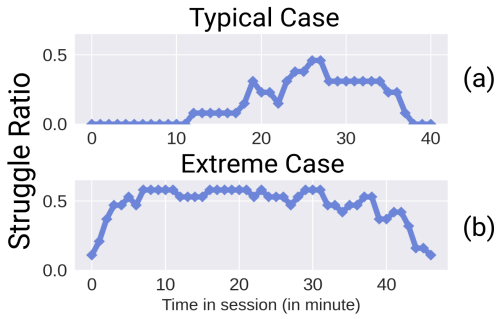

Figure 9. Struggle Ratio of a Typical (a) and Extreme (b) Case

Extreme Case. Among all sessions in our dataset, the per-minute struggle ratio rarely goes over 50%. By contrast, the extreme case session had more struggling students than non-struggling students in 27 out of 46 minutes, indicated by a struggle ratio of above 0.5, as shown in Figure 9 (b). This resulted in lower feasibility for Policy 1. The extreme case differs from the typical case in two aspects. First, unlike the typical case, almost as soon as the class began, students started wheel-spinning. Second, there were wheel-spinning students without potential partners in almost every minute of the session (indicated by red bars in Figure 10 (b)).

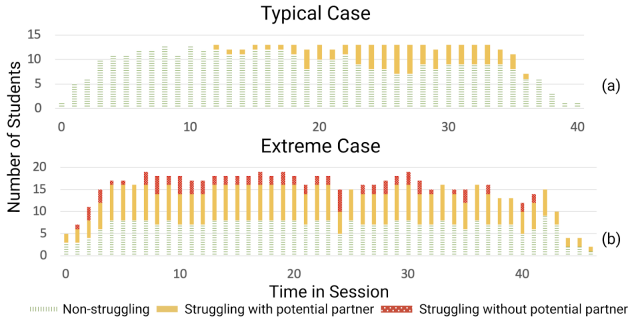

Figure 10. Policy 1 for a Typical (a) and an Extreme Case (b)

4.4.2 Policy 2

Same as in Policy 1, we chose this extreme case (Session 19, C3) as this policy had the worst performance on this session. This session also had the most different FI_1 trend. We simulated Policy 2 and calculated FI_1 for every minute in the two contrasting class sessions. Figure 11 showed the typical and extreme case, of how FI_1 changed when different PR rates were simulated. We plotted four different PR rates in the figure.

Typical Case. We saw two patterns in the Policy 2 simulation for the typical case. Firstly, the policy was typically robust in maintaining high feasibility with a non-zero (albeit low) PR rate. In Figure 11(a), lines with PR rate 0.1 and PR rate 0 completely overlapped. With these PR rates, there were no instances of struggle without a potential partner (i.e., feasibility was 1 across the whole session). Secondly, when the PR rate was high (0.5 or 0.8), FI_1 exhibited a sharp decrease, when there was an increase in student struggle. For instance, in Figure 9 (a) at minute 19, the struggle ratio increased from 0.07 to 0.23, as the number of wheel-spinning students went from 1 to 3. In Figure 11 (a) at the same time (t = 19 min), we saw a sharp decrease in FI_1 when the PR rate was 0.8.

Extreme Case. As shown in Figure 11 (b), the extreme case exhibited very different patterns compared to the typical case, mainly in three aspects. First, given it had a higher struggle ratio, even when there was no pairing restriction (i.e., PR rate = 0), we observed the FI_1 was not always 1 or even close to 1, as we saw in the typical case. Second, even a slight PR rate of 0.1 further worsened the policy feasibility and lowered the FI_1, unlike the typical case which showed resistance to a low PR rate. Third, if a class had a higher struggle ratio, the PR rate had a stronger effect on worsening FI, than for a session that had a lower struggle ratio. This effect was especially prominent when the PR rate was high (e.g., 0.5 or 0.8). This contrast means that the instructors may afford to set a higher PR rate without affecting the FI_1 too much, for a common session that has a moderate struggle ratio. However, the instructors may need to consider lowering the PR rate for a high-struggle session.

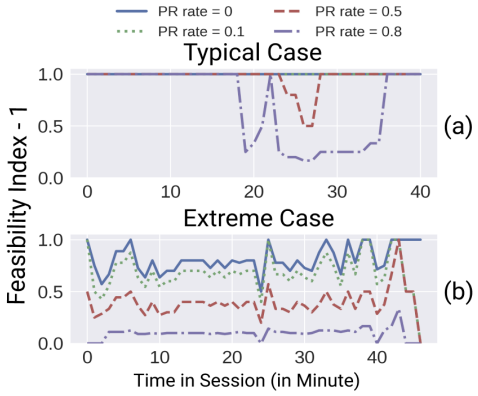

Figure 11. Policy 2 for a Typical (a) and Extreme (b) Case

4.4.3 Policy 3

In Figure 12 we present the results for Policy 3 simulation on two contrasting cases, plotting FI_2 for every step of the knowledge distance threshold for that session. The extreme case was chosen for having the most different FI_2 trend, from examining visualizations of FI_2 for all sessions involved.

Typical Case. As shown in Figure 12 (a), for the typical case, the FI_2 dropped gradually as the required KD threshold increased, which aligned with the overall simulation result. To pair students based on different knowledge (Policy 3), the instructors need to balance the required heterogeneity (i.e., higher knowledge distance threshold) and the desired paired ratio of the whole class. In this typical case, if a teacher selects a threshold of 0.5 or higher, none (0%) of students in the class session would be paired.

Extreme Case. As shown in Figure 12 (b), for the extreme case (Session 1, C5), while the downward trend was similar, we observed a more rapid decrease as compared to the typical case. Specifically, the FI_2 dropped to only 20% when the KD threshold was as low as 0.2, compared to 60% of FI_2 at the same KD threshold in the typical case. This comparison indicated that some class sessions were more heavily influenced by the parameter of the required knowledge distance threshold, and the effect may differ from session to session.

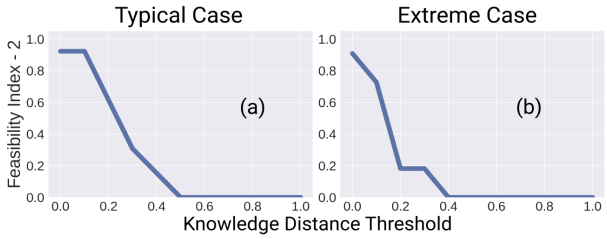

Figure 12. Policy 3 for a Typical (a) and Extreme (b) Case

4.4.4 Policy 4

From the previous analyses, we noted that Policy 4 performed reliably and similarly across classes, making it harder to select an extreme case or a worst case scenario. We selected a session where Policy 4 performed surprisingly well (Session 2, C3). In Figure 13 we visualized Policy 4 simulation on two contrasting cases, plotting FI_2 for every step of the knowledge distance ceiling γ for that session.

Typical Case. The tradeoff between knowledge homogeneity and policy feasibility was less prominent than under Policy 3. This means that instructors can afford to choose a stricter (i.e., lower) ceiling so students have a very small knowledge distance, and still achieve high feasibility (FI_2). For example, in Figure 12 (a), we saw that even if the instructor chooses a very strict threshold of $\gamma = 0.1$, nearly 95% of students were able to find a potential partner.

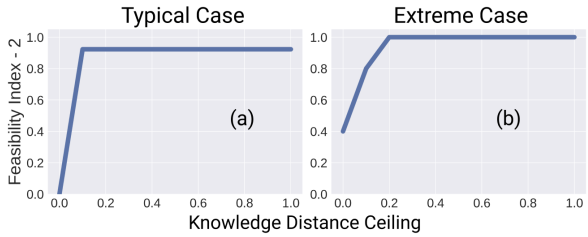

Figure 13. Policy 4 for a Typical (a) and Extreme (b) Case

Extreme Case. In Figure 13 (b), even when the KD ceiling was set to 0 (which means students must have the same level of mastery to be paired up), 40% of the students can still be paired, unlike typical cases where usually no two students have the exact same IK. Another noteworthy distinction is, while the typical case did not reach $FI_2 = 1$ with any ceiling of KD, the extreme case successfully paired all students ($FI_2 = 1$) with a relatively low ceiling of 0.2.

5. DISCUSSION

In line with previous LGF research [8], this work introduces four dynamic LGF policies contextualized in ITS and grounded in user research with K-12 teachers [10,30]. In this section, we discuss our main findings for research questions, grounded design recommendation for pairing orchestration tools, and future research direction for dynamic LGF.

5.1 Main Findings for Research Questions

Regarding the feasibility of the pairing policies (RQ1) and how the feasibility may depend on parameters of the pairing policies (RQ2), we found that *averaged across time and sessions,* it is generally feasible (93.6%) to team up struggling students with non-struggling students, the minute they struggle (Policy 1). This result remains true even when a high percentage of students is deemed ineligible for being teamed up with struggling students (Policy 2). Specifically, the average feasibility remains above 80% of struggling students across all sessions unless the pairing restriction rate is above 40%. However, as we see in our case study, there can be sessions and moments with high struggle ratios (hence, low feasibility) when using Policy 1. Relatedly, sessions with very high struggle seem more susceptible to the influence of the PR rate in Policy 2 than a typical session.

When pairing students based on whether their knowledge levels are different (Policy 3) or similar (Policy 4), the policy feasibility is highly dependent on the required KD. For Policy 3, there is a tradeoff between the desired heterogeneity (i.e., the knowledge distance threshold) and the policy's feasibility. This means instructors cannot set a high threshold for the KD if they want to pair most students. In Policy 4, the corresponding tradeoff (between homogeneity in knowledge and policy feasibility) is less prominent. Instructors may choose a stricter ceiling for students' similarity in knowledge levels and still achieve high policy feasibility. In the case study, we found that the policy feasibility in different sessions can be differently influenced by the required KD threshold or ceiling, depending on how closely clustered together students' IK is. For a given, fixed KD threshold (ceiling), a class of students closely clustered IK may result in higher feasibility for Policy 4 and lower feasibility for Policy 3. Presumably, the feasibility of these policies also depends on class size. For example, from our analysis, we hypothesize that for larger classes, the feasibility of pairing policies may change less drastically, when the policy parameters change or as the class progresses.

Regarding policy feasibility by class (RQ3), our results show no significant difference among classes for Policy 1 (Struggle with non-struggle) or Policy 4 (Knowledge similarity pairing). However, we observed significant differences among classes for Policy 2 (Pairing with restriction) and Policy 3 (Knowledge difference pairing). Although different classes have significantly different initial knowledge and struggle status, these differences in IK and struggle status are not always correlated with the feasibility of policies for that class. For example, classes that have different IK may not always have different feasibility for Policy 3 or 4.

5.2 Recommendations for Tool Design

The current study aims to inform the design of an orchestration tool that can help pair students dynamically. We aim to lessen teachers' orchestration load when managing *fluid social transitions*. Such a tool plays a key role in our vision for the smart classroom of the future, in which students alternate fluidly between individual and collaborative learning. Here, we highlight three design implications grounded in findings from the current work. These design implications may inform tools that aim to help teachers manage fluid social transitions, and ensure the feasibility of dynamic LGF policies. It may also offer inspirations, more broadly, for orchestration tools that aim to team up students in CSCL.

Firstly, technology could be used to automatically adjust the parameters used in LGF policies. Our study suggests that the four pairing policies studied provide a promising foundation for an orchestration tool, but greater flexibility is needed to deal with a wide range of circumstances than each individual policy provides. While some policies (e.g., Policies 1 and 2) explored in this study, have a good chance of working well during many class sessions, any given instantiation of a policy (with fixed parameter settings) does not fully deal with class variability and extreme cases. One way to compensate might be to have the tool automatically loosen policy parameters as needed. For example, the tool may gradually loosen the KD threshold or ceiling for policies 3 and 4, when it senses the pairing feasibility to be low.

Secondly, technology could use multiple LGF criteria in cascading fashion, to achieve high feasibility. Specifically, the tool may start out using the ideal pairing policies, and then iteratively try "more loose" criteria if the previous one fails to pair up all students. For example, the tool may first attempt to team up students based on struggle on specific KCs - a criterion that is more specific (and restrictive) than Policy 1, but one that could potentially be more effective for helping struggling students. If that fails, then it might pair up students based on their general struggle (Policy 1). If that fails again then the tool could try to pair students based on knowledge distance. The tool could also customize its pairing criteria such as using students characteristics that make the most sense in the given classroom context.

Lastly, technology could be used to recommend LGF policies or policy parameters with high feasibility to teachers. Instead of relying solely on the teachers to make pairing decisions, the tool

may adopt *SimPairing* to automatically calculate and maximize policies' feasibility based on classroom contexts and recommend them to teachers. For example, if the tool determined, using historical transaction data, that students in a given class have consistently low struggling ratios and fewer wheel-spinning students than non-wheel-spinning ones, it may advise that teachers adopt pairing Policy 1 as it has high feasibility. In addition, our findings open up the potential for the tool to help teachers make informed decisions about parameter configuration, by notifying them of expected feasibility. For example, if a teacher severely restricts the acceptable pairings, the tool could alert teachers of the low feasibility of the pairing policy, and ask if the teacher might want to loosen the restrictions. Running such simulations and providing notification can inform the teachers about the outcomes of the policy feasibility in their classroom, prior to implementing them. This may prevent teachers from choosing policy configurations that are misaligned with their goals.

5.3 Future Research Directions of Dynamic LGF

Building on our investigation, we outline four potential directions for how future research could further explore dynamic LGF policies.

Firstly, in addition to inter-session pairing based on knowledge distance (Policies 3 and 4), future work could explore *intra-session grouping based on knowledge level,* which allows forming pairings during the learning process. Researchers should explore how it would differ from our inter-session grouping and which approach better supports teachers' needs.

Secondly, the current policies identify the students to be wheel-spinning if they struggle on *any* of the KCs. Future work could explore whether teachers prefer to pair students based on their KC-specific struggle status. For example, to help a student struggling on the KC *combine constant terms* in equation solving, teachers may prefer to find a partner who has already mastered the same KC, or at minimum is not struggling on the same KC; they may (or may not) might find it acceptable, if the partner is struggling on another KC, e.g., *divide by variable coefficient*. Relatedly, teaming up students who are both struggling, but struggling on different knowledge components, may have benefits. Such a pair of students may have complementary knowledge and strength, and may help each other get unstuck and stop wheel-spinning. Such pairing criterion opens up good opportunities for tutor-tutee role-switching and mutual peer tutoring.

Thirdly, analogous to pairing based on KC-specific struggle status, instead of using the mean of students' mastery on different KCs to represent their knowledge, future work could explore to what extent *pairing students based on KC-specific knowledge distance* can be more effective, feasible, or preferable for teachers. KC-specific knowledge pairing might be useful for Policy 3, if teachers want two students who have very different skill levels on one specific KC so that the one with higher mastery on that KC can tutor the one with lower mastery.

Lastly, in addition to knowledge level and struggle status, which this work investigated for dynamic grouping, future work can investigate *other student characteristics* (e.g., history of collaborative episodes, preferences for working individually or collaboratively) or other *sources for knowledge level* (e.g., exams or quizzes score) for dynamic pairing. It may also be especially promising to further study pairing based on dynamic student behaviors that can be detected real-time by ITS from interaction data, to allow fluid social transitions and dynamic pairing.

5.4 Limitations

There is uncertainty in the *SimPairing* process in that we do not have a good way of estimating how long any given collaborative episode will last. Thus, *SimPairing* does not simulate students' being unavailable for pairing while they are working collaboratively, until they finish the collaborative episode. There is some reason to think that the resulting inaccuracy in the feasibility results is not severe, as argued, but we do not have a good way of investigating that issue in depth. Additionally, feasibility of pairing policies, while important, is just one piece of the puzzle. It is important, as well, to understand if students *learn better* with these pairing policies (effectiveness). Future research should validate these pairing policies in classroom studies, testing both their effectiveness and feasibility.

6. CONCLUSION

We study the feasibility of pairing policies in the context of ITS, to inform the design of a tool for orchestrating fluid transitions between individual and collaborative learning. Our findings show that on average, dynamically pairing students based on their in-the-moment wheel-spinning status results in good pairing feasibility for struggling students on average, even with moderate restrictions on the allowed pairings. We also found the trade-off between the *required knowledge distance* and *the policy feasibility,* is more prominent in heterogeneous grouping than in homogeneous grouping. However, any given instantiation of a policy (with fixed parameter settings) does not fully deal with class variability and extreme cases, as policies have different feasibility for different classes and sessions. This suggests optimization for policy feasibility (e.g., through gradually loosening parameters) or classroom customization need to be taken into consideration. Methodologically, this research extends previous work (e.g., Replay Enactments) that used authentic data and algorithms as design materials to augment designers' intuitions for designing future tools [27].

This work has several novel elements. First, using the *SimPairing* approach, our work explores the *feasibility* of LGF policies derived from user research with math teachers. In addition, to the best of our knowledge, this is the first study that considers students' in-the-moment wheel-spinning status in dynamic pairing policies. Finally, our work addresses a gap in the literature for dynamic intra-session LGF [1] and envisions how instructors and/or an orchestration tool will customize pairing policies and parameters to specific classroom contexts, which prior work argued to be especially helpful in the LGF process [1,8].

In sum, theoretically, this work bridges the literature gap on its investigation of the *feasibility* of user-centered dynamic pairing policies. Practically, we contribute grounded design directions for pairing orchestration tools, and *SimPairing* as an approach, to evaluate dynamic LGF policies, which may generalize to other online educational software that have transaction data.

7. ACKNOWLEDGEMENTS

This work was supported in part by Grant #1822861 from the National Science Foundation (NSF). Any opinions presented in this article are those of the authors and do not represent the views of the NSF. We thank Yanjin Long and Dr. Zach Branson for their help, and the anonymous reviewers for their feedback.

8. REFERENCES

[1] S. Amara, J. Macedo, F. Bendella, A. Santos, Group formation in mobile computer supported collaborative learning contexts: A systematic literature review. Journal of Educational Technology & Society. 19 (2016) 258–273.

[2] P. Dillenbourg. Over-scripting CSCL: The risks of blending collaborative learning with instructional design. P. A. Kirschner. Three worlds of CSCL. Can we support CSCL?, Heerlen, Open Universiteit Nederland. 61-91, 2002.

[3] X. Wang, M. Thompson, K. Yang, D. Roy, K.R. Koedinger, C.P. Rose, J. Reich, Practice-based teacher questioning strategy training with ELK: A role-playing simulation for eliciting learner knowledge. Proc. ACM Hum.-Comput. Interact. 5 (2021) 1–27.

[4] Y.-M. Huang, Y.-W. Liao, S.-H. Huang, H.-C. Chen, Jigsaw-based cooperative learning approach to improve learning outcomes for mobile situated learning. Journal of Educational Technology & Society. 17 (2014) 128–140.

[5] N.M. Webb, K.M. Nemer, A.W. Chizhik, B. Sugrue, Equity Issues in collaborative group assessment: group composition and performance. Am. Educ. Res. J. 35 (1998) 607–651.

[6] I. Srba, M. Bielikova, Dynamic group formation as an approach to collaborative learning support. IEEE Trans. Learn. Technol. 8 (2015) 173–186.

[7] A. Mujkanovic, D. Lowe, K. Willey, C. Guetl, Unsupervised learning algorithm for adaptive group formation: Collaborative learning support in remotely accessible laboratories. International Conference on Information Society (i-Society 2012), 50–57.

[8] N. Maqtary, A. Mohsen, K. Bechkoum, Group formation techniques in computer-supported collaborative learning: A systematic literature review. Technology, Knowledge and Learning. 24 (2019) 169–190.

[9] K. Holstein, B.M. McLaren, V. Aleven, Designing for complementarity: teacher and student needs for orchestration support in AI-Enhanced classrooms. Artificial Intelligence in Education. Springer International Publishing, (2019).157–171.

[10] J.K. Olsen, N. Rummel, V. Aleven, Designing for the co-orchestration of social transitions between individual, small-group and whole-class learning in the classroom. International Journal of Artificial Intelligence in Education. (2020) 24-56

[11] N. Ding, R.J. Bosker, E.G. Harskamp, Exploring gender and gender pairing in the knowledge elaboration processes of students using computer-supported collaborative learning. Comput. Educ. 56 (2011) 325–336.

[12] M.E. Lockheed, A.M. Harris, Cross-sex collaborative learning in elementary classrooms. American Educational Research Journal. 21 (1984) 275–294.

[13] D.A. Sandmire, P.F. Boyce, Pairing of opposite learning styles among allied health students: effects on collaborative performance, J. Allied Health. 33 (2004) 156–163.

[14] Y.-C. Kuo, H.-C. Chu, C.-H. Huang, A learning style-based grouping collaborative learning approach to improve EFL students' performance in English courses. Journal of Educational Technology & Society. 18 (2015) 284–298.

[15] E. Alfonseca, R.M. Carro, E. Martín, A. Ortigosa, P. Paredes, The impact of learning styles on student grouping for collaborative learning: a case study. User Model. User-Adapt Interact. 16 (2006) 377–401.

[16] Y. Taniguchi, Y. Gao, K. Kojima, S. Konomi, Evaluating learning style-based grouping strategies in real-world Collaborative Learning Environment, Distributed, Ambient and Pervasive Interactions: Technologies and Contexts. (2018) 227–239.

[17] P.-J. Chuang, M.-C. Chiang, C.-S. Yang, C.-W. Tsai, Social networks-based adaptive pairing strategy for cooperative learning. Journal of Educational Technology & Society. 15 (2012) 226–239.

[18] E.A. Day, W. Arthur, S.T. Bell, B.D. Edwards, W. Bennett, J.L. Mendoza, T.C. Tubré, Ability-based pairing strategies in the team-based training of a complex skill: Does the intelligence of your training partner matter? Intelligence. 33 (2005) 39–65.

[19] R. Niu, L. Jiang, Y. Deng, Effect of Proficiency Pairing on L2 Learners' Language Learning and Scaffolding in Collaborative Writing. The Asia-Pacific Education Researcher. 27 (2018) 187–195.

[20] J.A. Sutherland, K.J. Topping, Collaborative creative writing in eight−year−olds: Comparing cross−ability fixed role and same−ability reciprocal role pairing. J. Res. Read. 22 (1999) 154–179.

[21] N. Storch, A. Aldosari, Pairing learners in pair work activity, Language Teaching Research. 17 (2013) 31–48.

[22] L.S. Fuchs, D. Fuchs, C.L. Hamlett, K. Karns, High-achieving students' interactions and performance on complex mathematical tasks as a function of homogeneous and heterogeneous pairings. American Educational Research Journal. 35 (1998) 227–267.

[23] H.-W. Tien, Y.-S. Lin, Y.-C. Chang, C.-P. Chu, A genetic algorithm-based multiple characteristics grouping strategy for collaborative learning. International Conference on Web-Based Learning, Springer, (2013) 11–22.

[24] Y. Pang, F. Xiao, H. Wang, X. Xue, A Clustering-Based Grouping Model for Enhancing Collaborative Learning. 13th International Conference on Machine Learning and Applications. (2014) 562–567.

[25] B. Chen, G. Hwang, T. Lin, Impacts of a dynamic grouping strategy on students' learning effectiveness and experience value in an item bank−based collaborative practice system. Br. J. Educ. Technol. 51 (2020) 36–52.

[26] Y.-T. Lin, Y.-M. Huang, S.-C. Cheng, An automatic group composition system for composing collaborative learning groups using enhanced particle swarm optimization, Comput. Educ. 55 (2010) 1483–1493.

[27] K. Holstein, E. Harpstead, R. Gulotta, J. Forlizzi, Replay Enactments: Exploring possible futures through historical data. Proceedings of the 2020 ACM Designing Interactive Systems Conference, Association for Computing Machinery, New York, NY, USA, (2020)1607–1618.

[28] T. Nagashima, K. Yang, A. Bartel, E. Silla, N. Vest, M. Alibali, V. Aleven, Pedagogical Affordance Analysis: Leveraging teachers' pedagogical knowledge to elicit pedagogical affordances and constraints of instructional tools. International Society of the Learning Sciences (2020)

[29] N. Diana, M. Eagle, J. Stamper, Automatic peer tutor matching: Data-driven methods to enable new opportunities for help, (n.d.).

[30] V. Echeverria, K. Holstein, J. Huang, J. Sewall, N. Rummel, V. Aleven, Exploring human–AI control over dynamic transitions between individual and collaborative learning. In European Conference on Technology Enhanced Learning, Springer, Cham. (2020) 230–243.

[31] K.B. Yang, L. Lawrence, V. Echeverria, B. Guo, K.

Holstein, N. Rummel, V. Aleven. (Under Review). “I like student choice, program insights, but final say from the teacher”: Teachers’ Preferences regarding Human-AI Control in Dynamic Student Pairing. Manuscript submitted to EC-TEL 2021

[32] J.E. Beck, Y. Gong, Wheel-spinning: students who fail to master a skill, in: Artificial Intelligence in Education, Springer Berlin Heidelberg, 2013: 431–440.

[33] S. Kai, M.V. Almeda, R.S. Baker, C. Heffernan, N. Heffernan, Decision tree modeling of wheel-spinning and productive persistence in skill builders. Journal of Educational Data Mining. 10 (2018) 36–71.

[34] N. Matsuda, S. Chandrasekaran, J.C. Stamper, How quickly can wheel spinning be detected? Educational Data Mining, ERIC (2016) 607–608.

[35] C. Zhang, Y. Huang, J. Wang, D. Lu, W. Fang, J. Stamper, S. Fancsali, K. Holstein, V. Aleven, Early detection of wheel spinning: comparison across tutors, models, features, and operationalizations, International Educational Data Mining Society. (2019).

[36] T. Mu, A. Jetten, E. Brunskill, Towards suggesting actionable interventions for wheel-spinning Students. The 13th International Conference on Educational Data Mining, 183–193.

[37] K. Holstein, B.M. McLaren, V. Aleven, Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. Journal of Learning Analytics. 6 (2019) 27–52.

[38] N. Diana, M. Eagle, J. Stamper, S. Grover, M. Bienkowski, S. Basu, Peer tutor matching for introductory programming: Data-driven methods to enable new opportunities for help. International Society of the Learning Sciences. (2018)

[39] K.B. Yang, T. Nagashima, J. Yao, J.J. Williams, K. Holstein, V. Aleven. Can Crowds Customize Instructional Materials with Minimal Expert Guidance? Exploring Teacher-guided Crowdsourcing for Improving Hints in an AI-based Tutor. Proc. ACM Hum.-Comput. Interact. 5 (2021) 1–24.

[40] S. Ritter, J.R. Anderson, K.R. Koedinger, A. Corbett, Cognitive tutor: applied research in mathematics education, Psychon. Bull. Rev. 14 (2007) 249–255.

[41] Y. Long, V. Aleven, Supporting students’ self-regulated learning with an open learner model in a linear equation tutor, in: International Conference on Artificial Intelligence in Education, Springer, (2013) 219–228.

[42] K. Holstein, B.M. McLaren, V. Aleven, Student Learning Benefits of a Mixed-Reality Teacher Awareness Tool in AI-Enhanced Classrooms. Artificial Intelligence in Education, Springer International Publishing, (2018) 154–168.

[43] M. Waalkens, V. Aleven, N. Taatgen, Does supporting multiple student strategies lead to greater learning and motivation? Investigating a source of complexity in the architecture of intelligent tutoring systems, Computers & Education. 60 (2013) 159–171.

[44] A.T. Corbett, J.R. Anderson, Knowledge tracing: Modeling the acquisition of procedural knowledge, User Model. User-Adapt Interact. 4 (1995) 253–278.

[45] E. Walker, N. Rummel, K.R. Koedinger, Adaptive intelligent support to improve peer tutoring in algebra. International Journal of Artificial Intelligence in Education. 24 (2014) 33–61.

[46] E. Walker, N. Rummel, K.R. Koedinger, To Tutor the Tutor: Adaptive Domain Support for Peer Tutoring, Intelligent Tutoring Systems. 626–635.

[47] K.R. Koedinger, J. Stamper, P.F. Carvalho, Sharing and Reusing Data and Analytic Methods with LearnSphere, Hands-On, 2, 30p.

[48] M.V.Q. Almeda, When practice does not make perfect: Differentiating between productive and unproductive persistence. (2018). Doctoral dissertation, Columbia University.

[49] Y.-M. Huang, T.-T. Wu, A systematic approach for learner group composition utilizing U-learning portfolio, Educational Technology & Society, Vol. 14, (2011).

[50] D.W. Johnson, R.T. Johnson, Learning together and alone: Cooperative, competitive, and individualistic learning, 2nd ed, 2 (1987) 193.

[51] N.M. Webb, A.S. Palincsar, Group processes in the classroom, in: D.C. Berliner (Ed.), Handbook of Educational Psychology , (pp, Macmillan Library Reference Usa; London, England, New York, NY, US, (1996). 841–873.

[52] R. Hübscher, Assigning Students to Groups Using General and Context-Specific Criteria, IEEE Trans. Learn. Technol. 3 (2010) 178–189.