Algorithmic group formation and group work evaluation in a learning analytics-enhanced environment: Implementation study in a Japanese Junior High School

Changhao Liang[a], Rwitajit Majumdar[b], Yuta Nakamizo[a], Brendan Flanagan[b] and Hiroaki Ogata[b]

[a]Graduate School of Informatics, Kyoto University, Kyoto, Japan;
[b]Academic Center for Media and Computing Studies, Kyoto University, Kyoto, Japan

ARTICLE HISTORY
Compiled September 26, 2022

ABSTRACT
In-class group work activities are found to promote the interpersonal skills of learners. To support the teachers in facilitating such activities, we designed a learning analytics-enhanced technology framework, Group Learning Orchestration Based on Evidence (GLOBE) with data-driven approaches. In this study, we investigated how algorithmic group formation and group work evaluation systems were orchestrated in a Japanese junior high school throughout a series of collaborative learning activities. From the field implementation of 12 group formations, we validated the difference in the measured heterogeneity of the groups formed by the different algorithms to create homogeneous and heterogeneous groups compared to random grouping. Further, the peer rating and self-perception of the group work were compared for different contexts (comparative reading and idea exchange) conducted by grouping following different algorithms. We found that groups formed heterogeneously or homogeneously considering the learner model data performed better than random grouping. Specifically, students in groups created by the homogeneous algorithm received higher peer ratings and more positive self-perception of group work in the idea exchange group tasks. We did not find significant differences in the context of comparative reading. Additionally, we examined texts extracted from the peer feedback tags to reflect on the group work process and the usability aspects of the peer evaluation system. Along with empirical findings, this work presents a paradigm of continuous data-driven group learning support by incorporating the peer and teacher evaluation scores as an input to the subsequent algorithmic grouping.

KEYWORDS
Computer-Supported Collaborative Learning (CSCL) ; Learning analytics (LA) ; Group formation ; Peer evaluation ; Genetic algorithm

1. Introduction

Collaborative learning is progressively adapted in various pedagogical contexts. During collaborative learning, participants work together to share ideas, help each other or accomplish team goals (Dillenbourg, 1999), which benefits many of their soft skills development such as critical thinking, problem-solving, and interpersonal consultation that count in modern society (Stahl et al., 2006). Computer-supported collab-

CONTACT Changhao Liang. Email: liang.changhao.84c@st.kyoto-u.ac.jp

orative learning (CSCL) (Stahl et al., 2006) and learning analytics (LA) (Siemens, 2012) provides digital tools and data support, thus bringing immense opportunities to scaffold such activities with information technologies. Currently, many researchers focus on the implementation of LA tools during the orchestration phase of the group work (Rodríguez-Triana et al., 2015; Van Leeuwen et al., 2014), or in a synchronous digital learning environment (Van Leeuwen, 2015). However, valid support for group formation and evaluation phases in a classroom-based environment deserves further attention.

Meanwhile, obstacles to providing valid support to collaborative learning exist. In the face-to-face in-class group learning context of junior high schools, collaborative learning appears to happen in form of small-group learning (Gillies, 2003). Teachers should compose each group and align students appropriately according to different learning contexts (Urhahne et al., 2010). In terms of group formation, teachers tend to resort to random grouping or just pairing neighboring students owing to difficulties to do it in a real-time manner (Salihoun et al., 2017). Unlike online learning environments, students of traditional classrooms seldom use digital tools, which leads to a cold start problem for the lack of enough learning logs to create learner models (Brusilovsky et al., 2015) that can be used to allocate students based on their attributes. In the CSCL-supported context, there remains a chance that teachers would get overwhelmed if they do not know how to make the best of computer-supported tools for orchestration. In addition, to evaluate the performance of the in-group work, only the teacher's evaluation is not enough since one teacher cannot check what is happening in all groups during the group learning (Kasch et al., 2021). Thus such technology adoption barriers further demand teachers' efforts and distract from starting the classroom activities (Austin et al., 2010).

The limitations listed in actual practice prompted the introduction of technology support on classroom implementation of group work that can help the teacher create groups efficiently during the class and get more information on students' participation. Since such in-class group learning practice with digital systems and data support is still not studied to a great extent, this study implements systems of Group Learning Orchestration Based on Evidence (GLOBE) framework (Changhao et al., 2021a) in a classroom group work context. In this paper, we aim to show how the systems of GLOBE operate for in-class group work and depict a picture of a classroom group work implementation under learning analytics support.

2. Related works

2.1. Algorithmic group formation based on student model

Group formation is a starting point (Sadeghi and Kardan, 2016) and fundamental task (Wessner and Pfister, 2001) to achieve pedagogical goals of group learning. Collaborative learning with properly formed groups is found to outperform traditional teaching (Kyndt et al., 2013), while improperly selected group formation parameters may raise several problems that lead to failure (Wang, 2010). Since various issues such as group members' characteristics, the context of the group creation process and the techniques used to form the group (Maqtary et al., 2019) could affect the group learning processes. Janssen and Kirschner (2020) also mentioned that students' domain knowledge, collaboration skills, self-regulation skills, group size, and group experience should be considered when designing a group work activity. In Sánchez et al. (2021),

personalities such as gender, self-efficacy, and attitude were used but such one-off data was collected in one specific context. In the learning analytics-enhanced environment, there are student model data (Brusilovsky et al., 2015) that covers recorded learning behaviors on the learning management platforms (LMS), preferred learning styles, previous common working experiences, and so on (Bozic et al., 2008). The student model data depicts learning-relevant characteristics of learners that could be considered in group formation.

Meanwhile, researchers have pointed out that the optimal group composition varies from different pedagogical contexts with diverse goals and output (Manske et al., 2015). Though heterogeneity between group members and their resources is recommended according to Vygotsky (1980), homogeneous compositions in learning engagement patterns could avoid neglect and isolation of learners during group work (Salihoun et al., 2017). Also, Sanz-Martínez et al. (2019) found that homogeneity in learning engagement produced better quality in team assignment with more interactions and self-efficacy. Heterogeneous groups in intellectual abilities, gender, experiences, preferences, interests, personalities are implemented for better results by peer help (Knez et al., 2017). Kanika et al. (2022) showed heterogeneous composition in academic achievement on group performance as well, but for subjective perception, students were more satisfied with homogeneous group settings. Jensen and Lawson (2011) indicated that homogeneous grouping in terms of students' initial reasoning abilities performs better with more positive attitudes toward collaboration in the inquiry learning context, while in the didactic condition heterogeneous groups outperformed homogeneous groups.

To form groups with homogeneous or heterogeneous compositions, artificial intelligence (AI) algorithms can make use of the former attributes to generate groups based on optimization functions. For instance, evolution-based machine learning, which is flexible to different group compositions and matches the multidimensional input of student model attributes to provide a genetic solution of formed groups (Moreno et al., 2012; Sukstrienwong, 2017). According to Flanagan et al. (2021), multiple variables input are vectorized with the value of each dimension representing one input parameter. Then, the fitness value (F) indicating the distance of each vector, will determine homogeneous groups with a smaller F, or heterogeneous groups with a larger F. The fitness value can also play a role as the heterogeneity indicator of each group characterizing its internal composition. Besides, there are other methods such as clustering (Kanika et al., 2022; Maqtary et al., 2019) for homogeneous grouping, semantic analysis based on textual input (Erkens et al., 2019; Manske and Hoppe, 2016) and social network analysis (Sadeghi and Kardan, 2016; Yoshida et al., 2020). Given the open nature of the learner model attributes that Flanagan et al. (2021) approach can accommodate, we have adopted this method in the current version of the system. In summary, current systems on algorithmic group formation have provided abundant approaches for orchestrating optimized groups. These systems usually provide only one group formation strategy using fixed characteristics of learners for a specific context. Hence a more integrated system that enables flexible group formation approaches and selective student model data deserve further discussion.

2.2. Group learning activities: Classroom orchestration and practices

Group learning activities practiced in the classroom provides students with opportunities to acquire basic collaboration skills and valuable experience (Chowdhury et al., 2002) and can cover different learning contexts. Teachers may conduct group learning

for simple brainstorming to just share ideas (Chang and Yeh, 2021), or peer help-oriented activities that require collaborative knowledge construction (Fischer et al., 2002) such as collaborative reading (Toyokawa et al., 2021) and problem-solving tasks (Ouyang et al., 2021), or workshop with further project-based group cooperation and more complex activities (Zhang et al., 2011). In terms of the national language class, the former two contexts featuring in-class communication are more common, but the involvement level of the learner's previous knowledge is different. In a peer help-oriented context, more knowledge externalization and elicitation of task-related knowledge happen (Fischer et al., 2002), so that individuals can extend their knowledge by the distribution of knowledge resources within the group work according to Vygotsky's Zone of Proximal Development (ZPD) theory (Vygotsky, 1980).

To assess the classroom-implemented group learning process, several indicators reflect the performance of group learning activities. Arvaja et al. (2007) pointed out the perceived quality of experience is discernible as well, which shows the insight of group work participants in a direct way. In Japanese schools, the idea of "proactive, interactive and authentic learning" suggested by the national curriculum standards has been implemented as the goal of modern education as well as common evaluation criteria for pedagogical activities (Mikouchi et al., 2019). In case of lacking objective data like real-time communication records for process analysis, several studies inspect the classroom-implemented group learning process using quantitative methods such as observations (Ambreen, 2021) and focused-group interview for supplement (Šerić and Garbin Praničević, 2018).

2.3. Group work evaluation

The evaluation of group learning can not only provide a grade for the course but also improve group learning quality and give motivation during the process to promote individual learning (Forsell et al., 2020). The evaluation methods can be broadly divided into summative or formative assessment (Strijbos, 2010). Formative assessment is proved to be helpful to facilitate reflection and immediate correction (Aminu et al., 2021; Mentzer et al., 2017). Hence, in a data-rich environment, instant feedback and enriched group awareness information (Ollesch et al., 2019; Strauß and Rummel, 2021) were adopted to support the group work process.

However, one teacher cannot check what is happening in all groups during the group learning (Kasch et al., 2021; Van Leeuwen, 2015). There exist some solutions to this issue in the mobile learning context (Alvarez et al., 2021) by providing monitoring of the group work process for the teacher, classroom-based scenarios still need supplement from peer evaluation. Meanwhile, problems of social loafing and free riding (Strijbos, 2010) are prevalent that remain large obstacles to successful group learning activities.

Peer evaluation becomes imperative to alleviate teachers' workload and provide a real-time inspection across the group learning process (Willey and Gardner, 2010). The peer evaluation tools evolve from paper-based surveys to digital files and online platforms (Tharim et al., 2016), making the evaluation delivery process faster (Cleynen et al., 2020) with anonymity (Cheng and Warren, 1997), which can enable teachers to conduct the evaluation activities in a short time. Peer evaluation engagement also benefits to improving the students' soft skills such as critical thinking (Rohmah et al., 2021) and self-regulation (Meusen-Beekman et al., 2016). It can also enhance students' motivation and increase attendance as a facilitator (Chaloupskỳ et al., 2021). The quality of peer evaluation becomes a promising issue (Aminu et al., 2021). Cur-

rently, researchers aim to improve these peer evaluation skills using group awareness indicators from their learning logs (Kasch et al., 2021) or interactive peer evaluation platforms with backward feedback (Lin et al., 2021). From the learning analytics perspective, the re-use of these peer evaluation data is seldom discussed in the former peer evaluation platforms. Changhao et al. (2021b) went further on the role of peer evaluation in the data-driven ecosystem, and this study will serve as an empirical classroom implementation of that approach.

3. Group Learning Orchestration Based on Evidence (GLOBE) framework and its system components

Group Learning Orchestration Based on Evidence (GLOBE) provides a framework for group learning support with data-driven approaches in the learning analytics-enhanced environment (Changhao et al., 2021a). As illustrated in the figure 1, the data-driven workflow covers four phases: group formation, orchestration of group work, evaluation of group work, and reflection after group work. For this study, the algorithmic group formation system and the peer evaluation system instantiate the GLOBE framework as two organic components of a Learning Analytics Dashboard (Majumdar et al., 2019).

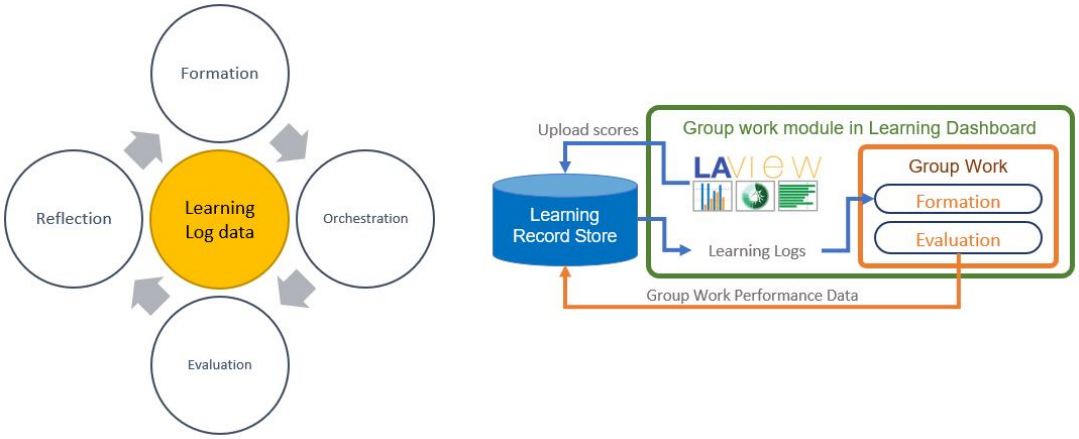

Figure 1. GLOBE framework and its two implement systems

3.1. Group formation module: algorithmic grouping using logs in student model

As for the group formation module, based on the former research that formed groups by the simple ranking of each characteristic (Changhao et al., 2021a), a genetic algorithm was applied to strengthen the flexibility to multiple data sources. To represent a group formation, one combination of students constructs a candidate individual (G) as a set of randomly-ordered students (s) partitioned by groups (Figure 2). For each student, there is a corresponding vector covering multiple characteristics of the student for the calculation of fitness value. These characteristics come from user model variables such as online reading logs, quiz scores from the LMS, and previous rating data from the peer evaluation module. Each dimension of a student vector is represented by a certain variable selected by the user. Figure 3 illustrated an example of metrics representation

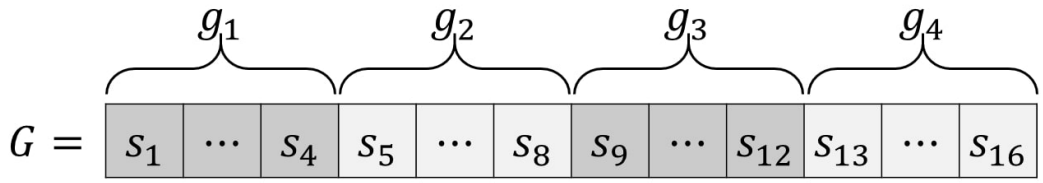

Figure 2. Representation of a candidate group formation as a vector of students divided into groups, illustrated by an example of 4 groups of 4 students (Flanagan et al., 2021)

where each student (s) is represented by a column vector with a characteristic (c) being represented as a dimension.

For the fitness estimation, the system uses the measure of squared differences. Adapted from the global optimization method of the original algorithm that concentrates on inter-group difference (Moreno et al., 2012), a local optimization strategy focusing on the intra-group difference of characteristics of members within each group (Flanagan et al., 2021) was used in this implementation. The Equation 1 shows the fitness calculation of each individual (G), where S is the number of students, C is the number of characteristics, N is the number of groups, and $\overline{x}_{j,g}$ is the average value of the characteristic j in the group g. The fitness value of one group formation (F) is the sum of all of the fitness values of each group (F_g). Employing the fitness value, we can determine homogeneous groups that have similar members and a small F, or heterogeneous groups that are made up of dissimilar group members shown by a large F. This fitness measure is used to cull undesirable candidates during the genetic algorithm iteration processes of breeding, crossover, and mutation (Flanagan et al., 2021) from the original candidate individual (G). Finally, it can select the best candidate (G) among all individuals with the largest or smallest F at the end.

$$F_g = \sum_{s=1}^{S} \sum_{j=1}^{C} \left(c_{j,s} - \overline{x}_{j,g}\right)^2, F = \sum_{g=1}^{N} F_g \tag{1}$$

The algorithms were used to create groups with different compositions. After creating groups, teachers can also check the group's homogeneity and the details of each attribute of the group members. Figure 4 serves as examples of a heterogeneous group and a homogeneous group formed by the system showing its F_g in the equation 1 as the squared differences within the group. This F_g value denotes the heterogeneity of the corresponding group. In this case, the group is created based on three student model variables: course score, teacher's ratings, and peer ratings. The course scores can be any academic performance score like quizzes, and the teacher's and peer ratings can be collected in the group work evaluation module introduced in the next subsection.

In the heterogeneous group, we can find a higher heterogeneity, where student 3 got zero in the course score, and student 4 received a lower peer rating in the past. Such extreme values can be attributed to the absence of previous group works. For the heterogeneous grouping algorithm, we can ensure that these student with missing previous data can be assigned to diverse groups so that those with previous group work experience can assist them. While when considering homogeneous grouping, teachers may exclude these students from the algorithm and manually assign groups for them

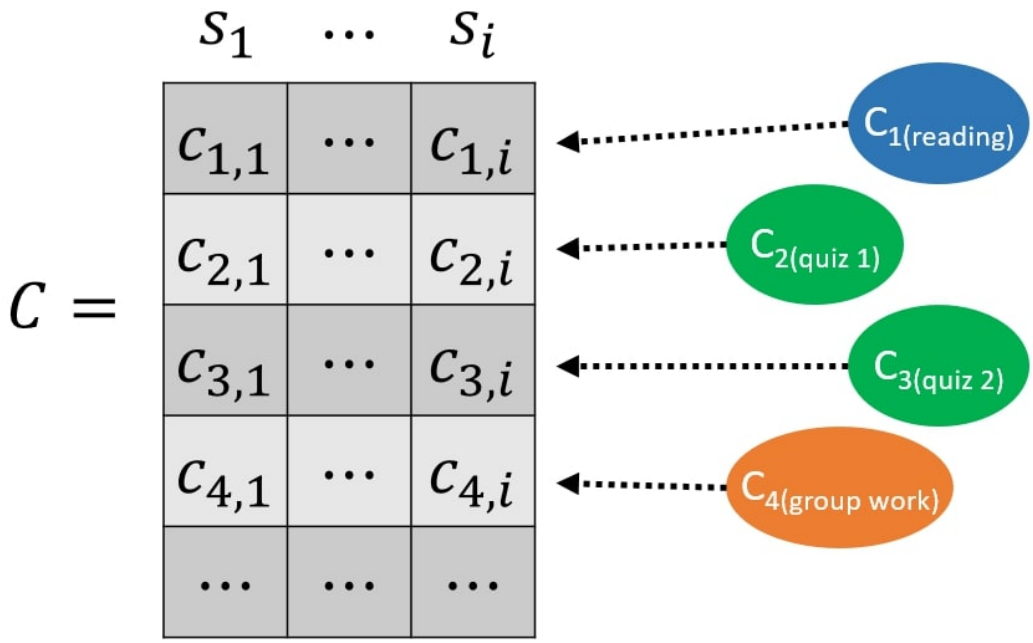

Figure 3. The mapping of student model variable values to the student characteristic representation matrix (Flanagan et al., 2021)

later.

Conversely, the squared heterogeneity of the homogeneous group is much lower where group members have closer scores. For random group formation, the members of each group are determined totally by random arrangement without any data intervention. Hence the heterogeneity of each group under random group formation remains unstable.

3.2. Group work evaluation module: System support for peers and teacher

The group work evaluation module provided the affordances to both teachers and peers to rate their evaluation of the group work. For the teacher's rating, the teacher can directly give ratings to each group in the group panel. In the peer evaluation module (Changhao et al., 2021b), group members can rate other individuals in their group or another group by just clicking the stars in the interface. They can also provide textual comments about the group learning as formative peer feedback. When students received feedback from peers, the comments will be visualized in the teacher's interface instantly. Once the ratings and comments are provided, the system shows them to the specific users with real-time ratings and textual feedback without association to the evaluator's name. The teacher can also set whether to show these ratings directly to the students as formative feedback or temporarily hide them and show them as a summative score later. Before the evaluation, the teacher can set the criteria of peer evaluation and the student can see each indicator of the criteria (for example subjectivity, communication, and perceived learning) as an independent column (see Figure 5).

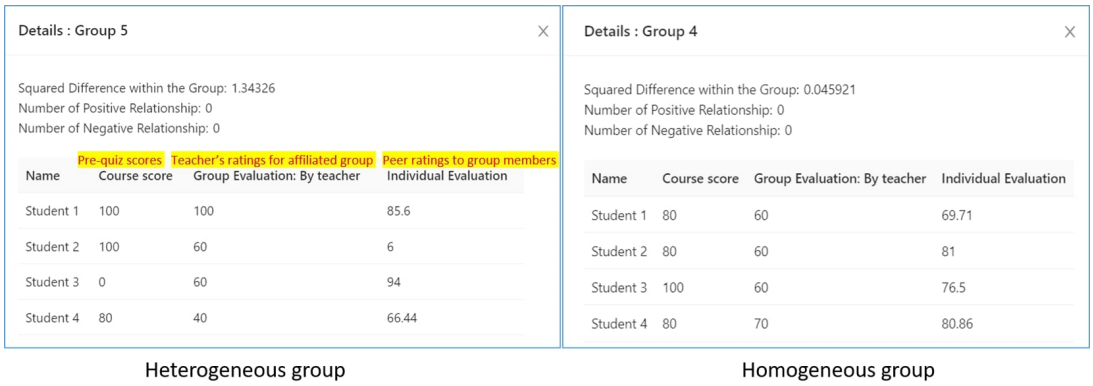

Figure 4. Example of the group formation details of a heterogeneous group and a homogeneous group (raw scores scaled to 0-100)

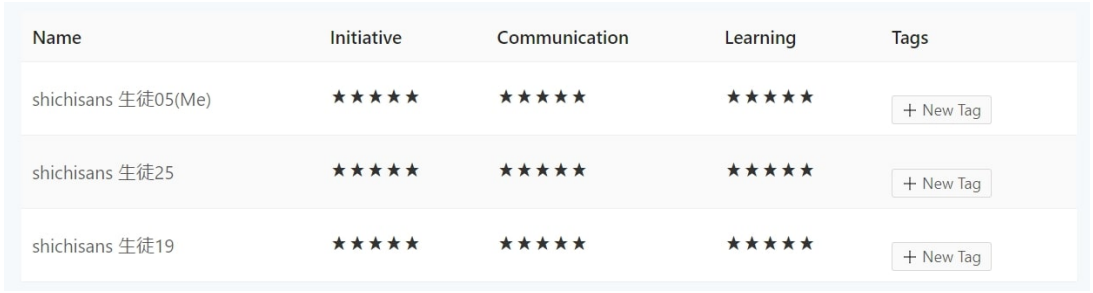

Figure 5. Interface of peer rating with three criteria set by the teacher

3.3. Supporting continuous data-driven group works using GLOBE

Figure 6 summarizes the continuous data-driven support throughout the two phases of GLOBE. As is depicted in the figure, the peer's evaluation together with the teacher's evaluation is logged into the learning record store as a part of the student model (orange circles). It can be reused as input to the algorithm in the following group formations (orange triangles). These inputs can also be used to identify students who may need special attention in the current group learning beforehand (Bukowski et al., 2017) in the detail panel of Figure 4. Additionally, to determine the reliability of each evaluator's peer ratings, the student model attributes used in the group formation phase can be capitalized on as performance indicators according to (Piech et al., 2013) to address the impact of biased peer scores. In other words, raters with higher scores in the group formation indicators will be modelled as high-reliability students and get a higher weight when calculating the scores for an individual using weighted average scores. Currently, this function has been visualized in the system (See Figure 7) showing both the raw score and the weighted score considering the reliability of each rater, and would be further investigated in the following research of this study. Also, simple randomized grouping followed by using the evaluation score for subsequent grouping provides a feasible solution to the cold start problem in data-driven research (van der Velde et al., 2021).

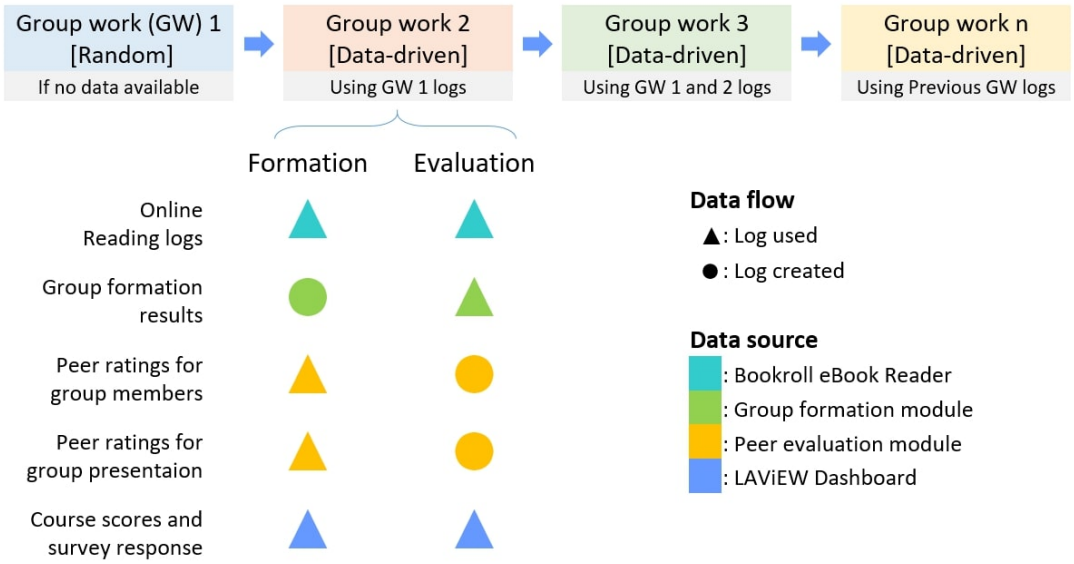

Figure 6. Example of a continuous data-driven support data flow under GLOBE

4. Research Method

4.1. Research Question

Based on the gap in the previous studies discussed in section 2 and the GLOBE framework presented in section 3, in this study we aim to investigate how the data-driven group formation together with group work evaluation systems work in an actual junior high school classroom context.

The specific research questions are stated as follows:

RQ1. Does data-driven algorithmic group formation create groups of different heterogeneity?

RQ2. What are the differences in students' peer ratings and self-perception of group work among groups created by different algorithms?

As for RQ2, we considered different algorithmic grouping conditions and divided RQ2 into two research questions:

RQ2.1 What are the differences of peer ratings and self-perception of group work between groups created by random arrangement and data-driven algorithmic group formation system for in-class group learning?

RQ2.2 For data-driven groups, what are the differences in peer ratings and self-perception of group work between groups created by the homogeneous and heterogeneous algorithm?

4.2. Study context and design

The study was implemented in native language classes of the second grade in a junior high school in Japan. The group learning activities focused on two contexts: idea exchange and comparative reading.

In the idea exchange context, students were expected to just share their opinions with group members, which is aimed to help them to get more inspiration and under-

Peer Evaluation : student 1

Details

Raw Score (Average): 4.33333
Weighted Score (Average): 4.209

Rater	Reliability	Raw Score
student 3	0.32	5.0
student 5	0.38	5.0
student 6	0.46	3.0

Figure 7. Example of a visualization of the weighted score considering the reliability of peer ratings

standings of the learning topic. Figure 8 shows the workflow of an actual idea exchange group learning (typical session 2 or 4).

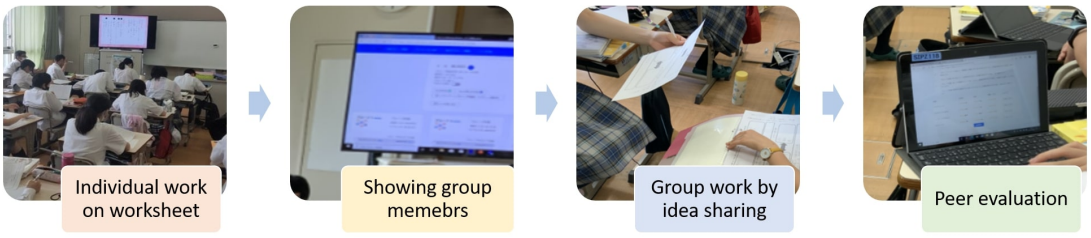

Figure 8. Idea exchange group learning: Classroom implementation workflow

In the comparative reading context (session 3), students were expected to find similarities and differences between two articles, which was aimed to help them to understand the topic from various perspectives and also practice their reading skills.

Such two contexts vary from the knowledge construction level since idea exchange may represent an activity with a low level of collaborative knowledge construction since sharing ideas with others does not require elaboration or critical discussion. While comparative reading may require higher-level collaborative knowledge construction in which the presented ideas are elaborated and critically discussed, and cognitive capabilities like reading skills are required as well.

A series of such activities across four sessions were conducted during the course topic of "the power of words". Each session took one class hour and was conducted sequentially within one week. For each session, the actual group work phase where stu-

dents discussed in small groups lasted 5 - 10 minutes. 12 group formations generated by random, homogeneous, and heterogeneous approaches were adopted in different classes (see Figure 9). Though the same sample of learners was compared in different conditions, they worked in different groups with different group heterogeneity, which is what we aimed to investigate in this study. Session 1 was an initiation to the system where students worked to understand the technology when participating in groups to think about a word while coming up with a name for a newborn baby. In session 2, students wrote down their opinion about the power of words in the worksheet individually by listing some daily words that they use. Then they shared their worksheet in groups and discussed them. In session 3, students were instructed to do comparative reading by working in groups. The output of this session tended to be more objective and reading skills-based compared to that of the previous idea exchange context. As for session 4, students first wrote a short composition about their impression of the power of words and then shared it with group members, which was similar to the idea exchange activity in session 2.

4.3. Participants

Participants were from grade 2 in a Japanese junior high school. 120 students (46 boys and 74 girls, with an average age of 14 years old) were selected by purposive sampling to be part of this study. They were distributed across 3 classes and were instructed by the same native language teacher. Each class had 40 students and there were 107 students (36 from Class A, 36 from Class B, and 35 from Class C) who participated in all sessions with some missed due to absence. Each Student with their parents had read and signed the consent form telling about privacy issues on personal data collection and usage.

4.4. Procedure

The procedure of the study across four sessions was summarized in Figure 9. For each group learning session, students were beforehand divided into groups using different group formation algorithms of the group formation system as is shown in the figure. We set the group size as four since it is easy for 4 students to sit around in the classroom with 4 neighboring tables, though some groups have only 3 members due to the absence issue. In the initiation activity of session 1 and idea exchange activity on session 2, students were combined by random arrangement without data intervention. Then, the algorithm used data from 2 to generate the groups of session 3 and the heterogeneity of these groups was measured by using the data at the end of session 3. It was the same with session 4 where data from both session 2 and 3 were utilized following the continuous data flow in Figure 6.

In session 1, a pre-test of reading comprehension related to the topic "the power of words" was conducted at the beginning, and a survey on attitude towards group learning (Cantwell and Andrews, 2002; Xethakis, 2018) was also incorporated after the class. From session 2, students were required to give peer ratings after the group learning. A 5-item self-perception of group work survey was also given at the end of class. After the class, the teacher gave ratings to each group depending on the activeness of communication as well.

When it comes to sessions 3 and 4, we used pre-test scores and previous ratings received by each student to generate homogeneous or heterogeneous groups for different

	Session 1	Session 2	Session 3		Session 4	
Context	Initiation	Idea exchange	Comparative reading		Idea exchange	
Class	All	All	Class A	Class B & C	Class A	Class B & C
Algorithm	Random	Random	Homogeneous	Heterogeneous	Heterogeneous	Homogeneous
Data Used			Pre-test score Teacher's ratings of class period 2 Peer ratings of class period 2		Pre-test score Teacher's ratings of class period 2 and 3 Peer ratings of class period 2 and 3	
Test	Pre-test				Post-test	
Evaluation		Teacher & Peer				
Survey	General attitude towards group learning (Xethakis, 2018)	Self-perception of group work (Drury et al., 2003)	Self-perception of group work		General attitude towards group learning & Self-perception of group work	

Figure 9. Procedure of the group learning experiment

classes. For each session executed in the same class, we employed different algorithms to control the learning effect caused by the order of the learning task. For session 3, students were grouped to do comparative reading tasks, where students in Class A were grouped homogeneously while Class B and C formed heterogeneous groups. Conversely, in session 4, Class A worked in heterogeneous groups and Class B and C worked in groups formed by the homogeneous algorithm for the idea exchange activity.

4.5. Instruments and data collected

We adopted Mixed Methods Research (Creswell et al., 2011) for data collection which covers both quantitative and qualitative data. The ratings from the teacher's and peer evaluation inputs are automatically collected in the data-driven evaluation system (Changhao et al., 2021b). The teacher walked around the classroom during the group learning and made some notes of the performance of each group. The teacher did the rating after the class since the scores are sensitive in Japanese high schools and he does not want students to see it directly. As for peer ratings, group members were asked to rate each other in three indicators: subjectivity, communication, and perceived learning, from the perspective of "proactive, interactive and authentic learning" suggested by the national curriculum standards of Japan. As is explained by Shiho (2021), "Subjectivity" indicates the motivation of the participation of the group work. "Communication" emphasizes student interaction through dialogue, which is measured by the activeness of speaking. "Perceived learning" refers to how much help you get from the member in the group work, which reflects the concept of "authentic learning" that focuses on the actual cognitive improvement. These three indicators have been implemented throughout daily pedagogical activities in Japanese schools since 2016 so that students were not alien to them (Mikouchi et al., 2019). A total of 506 evaluations from students were made in the system for the last three sessions. The teacher's evaluation scores were not considered in the data analysis of this research since we focused on students' evaluation this time, but these scores were used as the group formation input variables for sessions 3 and 4.

To measure the perception of group work, a 5-item self-perception of group work survey (see Table 1) was selected and adapted from the questionnaire of student per-

ceptions of group work in Drury et al. (2003) with a 5-point Likert-type scale from "strongly agree" to "strongly disagree". The Cronbach's alpha value of the survey was 0.901 in this study with relatively high reliability of the scales. To assume the homogeneity of three different classes, students took a pre-test of reading comprehension with 5 multiple choice questions in session 1 (e.g., "Read the article and choose which statement is right in the following answers."). A post-test with similar patterns was conducted in the end after finishing all 4 sessions. Meanwhile, a survey on the general attitude towards group learning based on Feelings Towards Group Work (FTGW) questionnaire (Cantwell and Andrews, 2002; Xethakis, 2018) composed of three constructs (Preference for Individual Learning (PIL), Preference for Group Learning (PGL), and Discomfort in Group Learning (DGL)), was also carried out in the initiation phase. The Cronbach's alpha values of FTGW in this study were 0.775 for PIL, 0.620 for PGL, and 0.546 for DGL which is similar to the related study (Xethakis, 2018).

Table 1. 5-item survey on the self-perception of group work (adapted from Drury et al. (2003))

No.	Item
1	I have had very positive experiences with group work.
2	The product of group work has been as good or better than I could produce as an individual.
3	We gave each member the opportunity to contribute.
4	I am a good player during the group work.
5	We work well as a group.

In addition, for the peer evaluation phase, we did random observations to find problems when students use the system in the actual classroom field. After the group activity, informal talks were conducted with the teacher and students after class.

4.6. Data analysis

Before analysis, we conducted tests to confirm the equivalence of groups by considering their academic performance and attitude to group learning. Table 2 shows the pre-test score proved to be of no significant difference in ANOVA so that we can consider each class performs similarly in academic performance. Meanwhile, it is also indicated that their post-test scores proved to be of insignificance, hence we can consider that the sequence of sessions does not affect the group work outcome.

To answer RQ1, we adopted ANOVA to examine the difference among the heterogeneity of groups created by different algorithms. To answer RQ2, firstly, we compared the students' ratings of groups formed by random arrangement and formed by data-driven algorithmic group formation system to answer RQ2.1. To control the issue of context difference, all group works under these comparisons were conducted in the idea exchange context. Then, as for RQ2.2, we went further to inspect the groups created by the homogeneous algorithm and heterogeneous algorithm in two group learning contexts. In this case, we divided different classes into different conditions and we have controlled the issue of inter-class difference as well as the sequence of sessions according to the former illustrations. For statistical examination, we took Mann-Whitney U tests since neither of the peer rating scores nor self-perception survey scores satisfied normal distribution according to Shapiro–Wilk test.

Table 2. ANOVA of pre-test score and attitude towards group learning survey

	Class	Mean	SD	N	F	η^2
Pre-test	A	4.128	1.490	39	0.039	0.0007
	B	4.216	1.134	37		
	C	4.167	1.464	36		
Post-test	A	3.821	0.451	39	1.372	0.025
	B	3.595	0.686	37		
	C	3.778	0.722	36		
PIL	A	12.333	3.578	36	0.672	0.013
	B	12.333	3.719	36		
	C	13.143	2.777	35		
PGL	A	23.611	3.055	36	0..63	0.009
	B	24.361	3.863	36		
	C	23.800	3.333	35		
DGL	A	10.222	2.542	36	0.627	0.012
	B	10.167	2.699	36		
	C	10.800	2.655	35		

5. Results

5.1. RQ1: Does data-driven algorithmic group formation create groups of different heterogeneity?

Table 3 lists the descriptive statistics of the heterogeneity of all groups created in this study under different group formation algorithms of the system measured by fitness values (Flanagan et al., 2021) introduced in section 3.1. For one group created by the homogeneous and heterogeneous algorithm, the fitness values are calculated automatically using the selected variables. For randomly-created groups, the fitness values are calculated manually using the values of the same variables. The results of ANOVA denote the significant difference ($F = 9.569$, $p < .001$, $\eta^2 = .18$) between groups created by three approaches, and Figure 10 shows the distribution. We can see groups created by the heterogeneous algorithm have higher heterogeneity values and those formed by the homogeneous algorithm have lower values. The groups formed by random arrangement are between those formed by two algorithms.

Table 3. Descriptive statistics and ANOVA of group heterogeneity under three group formation approaches

Algorithm	N	Mean	Min	Max	SD	F	η^2
Random	30	0.404	0.053	1.383	0.401	9.569***	0.18
Homogeneous	30	0.297	0.028	1.343	0.364		
Heterogeneous	30	0.687	0.137	1.361	0.422		

***$p < .001$.

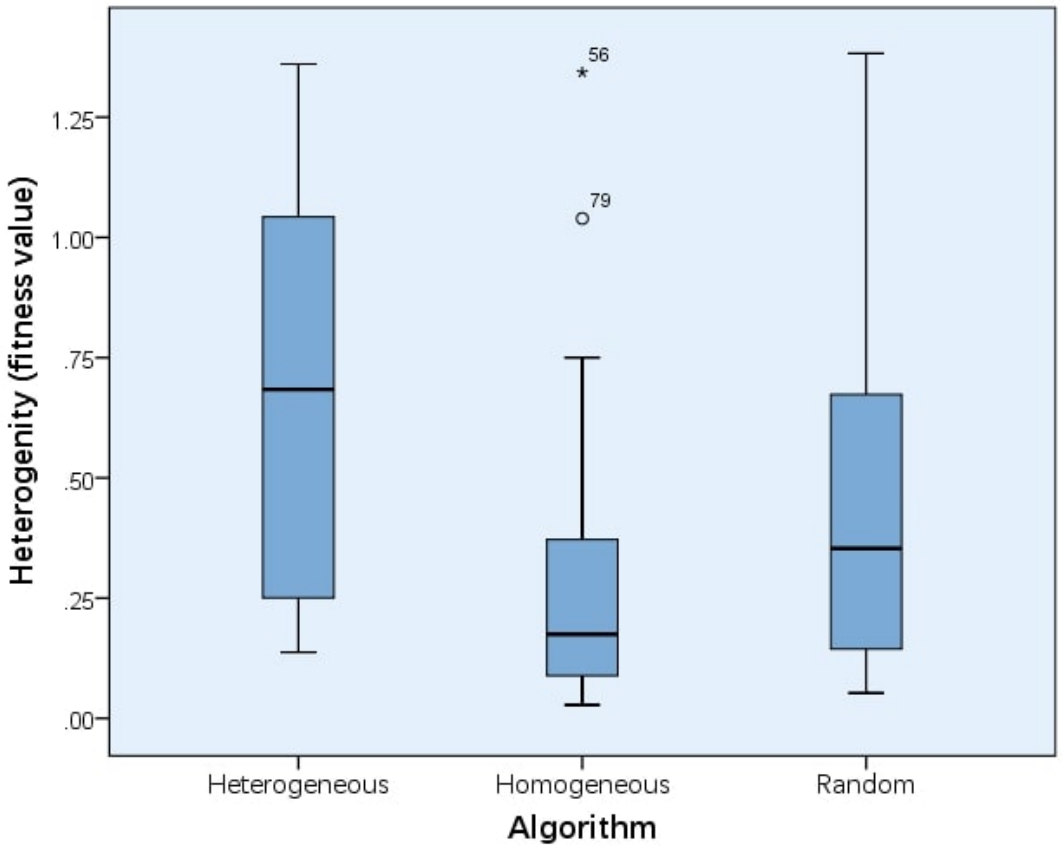

Figure 10. Box plot comparing heterogeneity of groups created by three approaches

Table 4. Post Hoc Comparisons of groups formed by different approaches

		Mean Difference	t	p_{tukey}
Heterogeneous	Homogeneous	0.390	4.234***	$< .001$
	Random	0.283	3.070**	0.008
Homogeneous	Random	-0.107	-1.164	0.478

Note. P-value adjusted for comparing a family of 3

Post-hoc tests found significant difference in heterogeneity between groups created by heterogeneous algorithm and homogeneous algorithm ($t = 4.234$, $p_{tukey} < .001$), and heterogeneous algorithm and random arrangement ($t = 3.070$, $p_{tukey} < .01$) (See Table 4).

5.2. RQ2: What are the differences in students' peer ratings and self-perception of group work among groups with different heterogeneity

5.2.1. Comparison of groups created by data-driven algorithmic group formation and random arrangement

Table 5 gives the overall result of statistical examinations with the green color indicating significance. Each comparison is independent since the sample is different due to different group compositions in each condition. As is indicated in the figure, groups formed by the homogeneous algorithm tend to have significantly higher peer rating scores as well as self-perception than random groups, and also perform better than groups formed by the heterogeneous algorithm in peer ratings. The specific results are discussed in the following subsections.

Table 5. Overall results of comparative studies of groups created by data-driven algorithmic group formation and random arrangement

Comparison of group composition	Sample of comparison[1]	Peer ratings[2]	Self-perception of group work
Heterogeneous (He.) v/s random (Ra.)	Class A session 4 (4-A) v/s Class A session 2 (2-A)	S: He. > Ra. C: He. > Ra. L: He. > Ra.	He. > Ra.
Homogeneous (Ho.) v/s random (Ra.)	Class B&C session 4 (4-B & 4-C) v/s Class B&C session 2 (2-B & 2-C)	S: Ho. > Ra. ** C: Ho. > Ra. * L: Ho. > Ra.***	Ho. > Ra. **

* p<0.05, ** p<0.01, *** p<0.001
1. The sample of each session is independent since the group composition changes in different session.
2. Consists of three sub-indicators: S – Subjectivity, C – Communication, L – Perceived learning

5.2.1.1. Groups under heterogeneous algorithm and random arrangement. As is shown in Table A1 and A2, groups with heterogeneous pre-test and past group learning performance scores got higher peer ratings in all three sub-indicators (subjectivity ($p = .520$), communication ($p = .445$), learning ($p = .051$)). The standard deviations of peer ratings and the self-perception survey are also smaller in groups formed by the heterogeneous algorithm. Students had a little bit higher score on the self-perception survey for groups formed by the heterogeneous algorithm as well ($p = .831$). However, all of these indicators do not show significant differences under the Mann-Whitney U tests.

5.2.1.2. Groups under homogeneous algorithm and random arrangement. As is shown in Table A3 and A4, groups with homogeneous pre-test and past group learning performance scores got higher peer ratings in all three sub-indicators. Also, they were more fulfilled in groups formed by homogeneous algorithm according to the self-perception survey of group work: the difference between two compositions on all peer rating indicators (subjectivity ($p = .003 < .01$, effect size $= .291$), communication ($p = .037 < .05$, effect size $= .202$), learning ($p < .001$, effect size $= .354$)) and self-perception survey ($p = .003 < .01$, Cohen's $D = .305$) showed statistical significance.

5.2.2. Comparison of groups created by heterogeneous and homogeneous algorithms

Since the contexts of session 3 and session 4 are different in the knowledge construction level as is mentioned in section 4.2. We will inspect the results in two different contexts. Table 6 summaries the comparisons under two different contexts.

Table 6. Overall results of comparative studies of groups created by heterogeneous and homogeneous algorithms

Group learning context	Sample of comparison	Peer ratings[1]	Self-perception of group work
Idea exchange	Class A session 4 (4-A) v/s Class B&C session 4 (4-B & 4-C)	S: Ho. > He. C: Ho. > He. L: Ho. > He. *	Ho. > He.
Comparative reading	Class B&C session 3 (3-B & 3-C) v/s Class A session 3 (3-A)	S: He. > Ho. C: Ho. > He. L: He. > Ho.	He. > Ho.

* $p<0.05$
1. Consists of three sub-indicators: S – Subjectivity, C – Communication, L – Perceived learning

5.2.2.1. Idea exchange context. As is indicated in Table A7 and A8, in the idea exchange context (session 4), groups formed by homogeneous algorithm got higher scores from both peer ratings (subjectivity ($p = .119$), communication ($p = .097$), learning ($p < .042$, effect size = -0.223)). They also had more positive persecutions on the group learning experience in the groups formed by the homogeneous algorithm according to the survey ($p = .108$). Only the perceived learning indicator showed significance in the Mann-Whitney U test.

5.2.2.2. Comparative reading context. As is shown in Table A5 and A6, in comparative reading context (session 3), groups formed by heterogeneous algorithm get higher scores for subjectivity indicator ($p = .662$), perceived learning indicator ($p = .635$) with less standard deviations, while groups formed by homogeneous algorithm got higher ratings in the communication indicator of the peer rating ($p = .293$). In addition, students had more positive perceptions of the group learning experience in heterogeneous groups ($p = .297$) according to the survey. However, none of these indicators implied a significant difference and there is almost no difference in terms of the subjectivity and communication indicators.

6. Discussion

6.1. Impact of algorithmic group formation system on actual group heterogeneity

For RQ1, the study shows the effectiveness of the group formation system under the GLOBE framework using a genetic algorithm to form groups with homogeneous or heterogeneous compositions. It contributes to the CSCL research area with a new

indicator, group heterogeneity, derived from the concept of fitness value in the genetic algorithm (Moreno et al., 2012), which can reflect how the group members are different or similar in the selected characteristics. In turn, it can be used to explain the findings of the difference of performance and outcome in the actual group work among groups with different heterogeneity values. According to the result of Figure 10, the system could successfully create groups with different with-in group differences according to the selected algorithm.

Studies on algorithmic group formation systems tend to focus only on heterogeneous groups (Haq et al., 2021) or homogeneous groups in specific characteristics of group members (Moreno et al., 2012; Sánchez et al., 2021). Compared to these researches, this system deliveries the flexibility that enables users to choose the algorithm as well as self-defined input variables, thus indicating potential implications on diverse learning contexts. The study extends the basic idea of using the genetic algorithm to form optimized groups (Moreno et al., 2012) in the educational context, and implement the algorithmic group formation method in (Flanagan et al., 2021) in a real classroom and conducted in-class group learning activities using the groups with different heterogeneity in real student model data from the digital platforms.

We can also see for groups created by the heterogeneous algorithm looks scattered, which means the heterogeneity of some groups formed by the heterogeneous algorithm was not high enough. Also, there are individual outliers with the values of its heterogeneity far from the corresponding algorithm. Though the average heterogeneity of data-driven groups is significantly different from that of random groups, such undesirable distributions may be a factor that causes the insignificance of the difference of peer ratings and self-perception between groups formed by the heterogeneous algorithm and random arrangement. To solve this issue, hence the coefficient of iteration times and the number of the evolution population need to be tuned for higher accuracy with the distribution of groups more centralized. Meanwhile, we measure the difference of each characteristic within group members using squared difference as (Moreno et al., 2012) did, which could get misleading when there are more outliers (Motulsky and Brown, 2006). For further improvement of the algorithm, more distance measures such as Cityblock, Euclidean, and Chebyshev should be considered as is suggested by Flanagan et al. (2021).

6.2. Connection of group heterogeneity with student-perceived group work outcome

As for RQ2, we addressed the comparison of peer ratings and self-perception of group work of groups created by different approaches. In terms of the comparison of random groups and data-driven groups for RQ2.1, our experiment showed that generally data-driven groups formed by the group formation system perform better than random ones in peer ratings and self-perception, especially homogeneous groups, while for heterogeneous groups the difference was very small in some indicators. The results agreed with our former research in primary school class, where we found groups formed by the system had higher engagement and positive affections than teacher-formed groups (Changhao et al., 2021a). Figure 10 indicates that the group heterogeneity may correlate to students' ratings and perceptions.

To further explain whether the heterogeneity of groups made such a difference in our findings, we inspected the group heterogeneity in each session. Based on the Mann-Whitney U test, we found that for the comparison of the random group session (2-A)

and heterogeneous group (4-A), the average heterogeneity of the 4-A session is higher. Though it does not reach a significant level in the Mann-Whitney U test ($p = .436$). For the comparison of the random group session (2-B & 2-C) and heterogeneous group (4-B & 4-C), the average heterogeneity of 4-B and 4-C sessions is lower ($p=.043 < .05$, effect size = .375), which indicated that students with common characteristics in the student model tended to be grouped together. Since this finding is consistent with the results of RQ2.1, it could give a possible explanation of our findings.

For the comparison of homogeneous and heterogeneous groups for RQ2.2, we inspected the effects in two different contexts. Results denote that groups formed by homogeneous algorithm perform better in all the indicators of the peer ratings for idea exchange context though only the perceived learning indicator reached the significant level. For comparative reading tasks, there was almost no difference in compared samples. The former result supports Sanz-Martínez et al. (2019) and manifests the impact of homogeneous composition on group interaction and self-perception of group learning experience in the idea exchange context. This result also agrees with group learning in online context (Abou-Khalil and Ogata, 2021) where groups formed by homogeneous algorithm enable learning achievement of low-engagement students and the self-perception of high-engagement students.

In terms of the latter result, related researches found that groups with heterogeneous knowledge levels adapt to peer help activities for better achievement (Kanika et al., 2022; Zamani, 2016) since there exists an imbalance of reading capabilities among students that level a foundation for peer help according to Zone of Proximal Development (ZPD) theory (Vygotsky, 1980). However, in this research, the difference is very small with a pretty low effect size. This may be caused by the variables we choose for group formation. In this study, we only used pre-test scores and ratings of former sessions as group formation input. The heterogeneity in such limited indicators may not reflect the diversity of the previous knowledge and skills of students. More student model variables and social-emotional characteristics such as personality traits (Sánchez et al., 2021) should be covered in the future study.

Meanwhile, other factors may contribute to the observed small, non-significant differences between the homogeneous and heterogeneous groups such as the differences between the Classes A, B, and C, and the sequencing of the sessions though we aimed to control them using test scores and group work attitude questionnaire. The imbalance of the samples in comparisons of homogeneous groups and heterogeneous groups could also affect the statistical results. As is shown in table A5 to A8, Classes B and C had higher ratings and self-perception scores in every context, which should be caused by their larger sample size. Based on the specific population and environment of Japanese junior high schools in the study, external validity needs to be further inspected under context in different cultures.

In addition, we have to admit that the peer ratings and self-perception can not perfectly reflect the whole picture of the group work process, and the impact of the heterogeneity on more group work outcome indicators such as the content of group discussion and ratings following more strict rubrics should be considered. In the following research design, we should collect more objective indicators. For example, the expert grading of the worksheet proceedings in each session.

Since it was found in earlier research that not every evaluator is capable of rating fairly (Carless and Boud, 2018), in such a scenario the reliability of peer evaluation might have been low because of novice raters who were doing such peer rating most for the first time. Also, there could be a tendency that students in a well-performed group tend to give higher scores to their group mates, while they might get harsher

in their peer evaluations if the group is failing to meet the course standard. In our observation we found a few students talking while doing the peer ratings though we do not know whether it was about the ratings. However, most of the students finished the peer grading individually and got used to the system in the latter sessions. Such bias towards peer evaluation also needs to be distinguished and accounted for when aggregating the scores. In this study, the peer rating scores of each student were calculated by the average of ratings from all evaluators regardless of the reliability. One possible approach to improve would be to assign a weight to different evaluators when integrating peer evaluation scores of each student. The weight can be estimated from their other student model attributes according to Piech et al. (2013) and used to correct the raters with low reliability by assigning them lower weight when integrating the peer ratings of each student. Another way of estimating the reliability of raters includes the correlation to the teacher's rating (Lin et al., 2021) and backward evaluation (Misiejuk and Wasson, 2021), which would be considered in the future system development.

6.3. Dynamics and potentials of peer evaluation system usage

In the peer evaluation phase, in light of the observation purpose of discovering potential problems, we found several obstacles for the first time to use the system. These obstacles were solved when students got used to the procedure in the latter sessions. The finding indicates that students can finish the peer evaluation task for in-class group work in a short time manner with the help of the digital system after their initial exposure and was encouraging for future implementations. Students as users also provided some suggestions on the user interface of the system after their use during the study sessions. Feedback included making the rating are bigger and more colorful. They also suggested the evaluation criteria could be more specific though they were familiar with the three indicators of the rating criteria out of their understandings. In the future, we shall take more efforts to elaborate the criteria in detail as suggested in Gueldenzoph and May (2002), and do some training sessions before the experiment.

Meanwhile, the textual comments from students help to figure out the group work process and some explanations for irregular patterns of their peer ratings. For instance, the comments from one group member of Session 2 Class B Group 4 disclosed the reason why student B23 kept talking with another group: "Student B23 spoke ill to student B19", and it could provide cues for the teacher's intervention if the teacher checked the comments in time. Another comment saying about the invalid talking that was unrelated to the topic uncovered further details behind the talking behaviors we observed, which would be hard for the teacher to detect. These comments from the peer evaluation system record the process of group work and enclose incentives to ratings with low reliability and in turn, make a breach to improve students' appreciative critical abilities (Rohmah et al., 2021). The finding supports the idea that peer evaluation can provide more information that is prone to be neglected by the teacher (Van Leeuwen, 2015). Students can receive more sufficient and instant feedback from peers than the teacher, which can be a central part of the learning process (Liu and Carless, 2006). Since we do not have much process data in this research, in the future research design, it's recommended that the teacher encourage students to use the comment function to record the process of the group work, and we could also encourage them to give more constructive comments on how the evaluatee could have performed better (Aminu et al., 2021) in the following experiments. Then, social network analysis and content analysis should be also adopted to construct peer evaluation networks and

discover further characteristics as was pointed out in Wang et al. (2020).

6.4. Challenges and implications of LA-enhanced group work orchestration

There were challenges when we planned the group work activities with the teacher since the teacher was unfamiliar with the LA-enhanced systems in the traditional middle school classroom, and the lack of student model data limited the power of the data-driven systems. To solve these problems, we designed a feasible workflow shown in Figure 6 for the teacher in this study based on the learning context of Japanese junior high school and showed the possibility to conduct the GLOBE framework in the face-to-face in-class group learning context. We started from the traditional group work in the initialization phase and gradually activated the continuous data collection and usage flow by generating data using the group work evaluation system within three sessions of group work.

In terms of the algorithm and the data-driven system, we underscored the heterogeneity of groups herein, while the selection of appropriate variables to consider in the algorithm was less discussed. As is suggested by Janssen and Kirschner (2020), multiple issues can affect the group work as antecedent attributes including not only group level characteristics like heterogeneity, but also individual characteristics that should be indicative or appropriate to indicate performance heterogeneity. In other words, what is heterogeneous is of equal importance in an education context (Cress, 2008). Therefore, we must admit that the current study could provide only part of the answer to this problem, and finding the right set of variables to accurately describe heterogeneity in a particular context remains a challenging task.

As for pedagogical implications, it provided a low threshold for the teacher to adapt the workflow thus promoting the use of a data-driven environment in actual class activities. Though we only used a few student model data in this implementation, it disclosed the opportunity for the LA-enhanced group work orchestration in a classroom-based context following the continuous data flow. Following the GLOBE framework, similar group work implementation could be done with this workflow in other in-class learning contexts such as math problem-solving (Changhao et al., 2021a) and English reading (Toyokawa et al., 2021).

As for technical implications, though there have been studies discussing the different impacts of groups formed by the homogeneous or heterogeneous algorithms in actual group work, this study contributes to digitizing this issue by introducing the heterogeneity value of each group which derives from the fitness value in genetic algorithm (Flanagan et al., 2021). Hence we provide a new perspective to explore details on how group heterogeneity makes a difference in group work as a meaningful step to the right direction. Under the affordance of this data-driven environment, further studies can be easily implemented to explore predictive variables for group formation. For instance, by investigating the specific student model variables for group formation we can figure out which characteristics the heterogeneity is more important to affect the group work process and outcome.

7. Conclusion and Future work

In conclusion, the study elaborates the features and practical implications of the algorithmic group formation and evaluation system of the GLOBE framework. Our

implementation also provided an example of how to start with no existing learning logs in student model initially and then incorporate the group work evaluations data cyclically for eventual group formation (Figure 6).

The empirical research conducted in this study illustrates an instructive practice of data-driven group learning implementation under the GLOBE framework. The impact of the algorithm-based group formation system to create groups with different heterogeneity, and inspects what difference does the group heterogeneity makes on the students' perceived group learning outcome. Results found that data-driven groups created by algorithmic group formation system received higher peer ratings than groups formed by random arrangement, and groups formed by homogeneous algorithm significantly more in idea exchange tasks. Based on this implementation, we enlightened the opportunity of the LA-enhanced group work orchestration in future classroom-based practice.

In future work, we aim to inspect groups created by more student model variables and explore the heterogeneity of which characteristics of group members cause the difference in group work performance. With the accumulation of data from various group learning contexts, automatized suggestions of optimal input variables to the teacher depending on the identified context could become possible. Also, group learning in the online environment with abundant learning logs in the student model deserves our further exploration. How to enhance peer evaluation reliability and cultivate critical abilities of students utilizing student model data and existing data-driven systems turns out to be another topic to explore.

Funding

This work was partly supported by JSPS KAKENHI 16H06304 and 20K20131 and 20H01722 and NEDO JPNP20006 and JPNP18013 and SPIRITS 2020 of Kyoto University and Support for Pioneering Research Initiated by the Next Generation program operated by the Japan Science and Technology Agency (JST) JPMJSP2110.

Ethics approval

This research work includes human participants and their participation has been vetted and approved by the Experiment Ethics Committee of Kyoto University.

References

Abou-Khalil, V. and Ogata, H. (2021). Homogeneous student engagement: A strategy for group formation during online learning. In *International Conference on Collaboration Technologies and Social Computing* (pp. 85–92). Springer.

Alvarez, C., Zurita, G., Carvallo, A., Ramírez, P., Bravo, E., and Baloian, N. (2021). Automatic content analysis of student moral discourse in a collaborative learning activity. In *International Conference on Collaboration Technologies and Social Computing* (pp. 3–19). Springer.

Ambreen, S. (2021). Children's perspectives toward ability-based group work in a primary classroom. *Journal of Research in Childhood Education*, 35(4), 651–665.

Aminu, N., Hamdan, M., and Russell, C. (2021). Accuracy of self-evaluation in a peer-learning environment: An analysis of a group learning model. *SN Social Sciences*, 1(7), 1–17.

Arvaja, M., Salovaara, H., Häkkinen, P., and Järvelä, S. (2007). Combining individual and group-level perspectives for studying collaborative knowledge construction in context. *Learning and Instruction*, 17(4), 448–459.

Austin, R., Smyth, J., Rickard, A., Quirk-Bolt, N., and Metcalfe, N. (2010). Collaborative digital learning in schools: Teacher perceptions of purpose and effectiveness. *Technology, Pedagogy and Education*, 19(3), 327–343.

Bozic, N. H., Mornar, V., and Boticki, I. (2008). Introducing adaptivity and collaborative support into a web-based lms. *Computing and informatics*, 27(4), 639–659.

Brusilovsky, P., Somyürek, S., Guerra, J., Hosseini, R., Zadorozhny, V., and Durlach, P. J. (2015). Open social student modeling for personalized learning. *IEEE Transactions on Emerging Topics in Computing*, 4(3), 450–461.

Bukowski, W. M., Castellanos, M., and Persram, R. J. (2017). The current status of peer assessment techniques and sociometric methods. *New directions for child and adolescent development*, 2017(157), 75–82.

Cantwell, R. H. and Andrews, B. (2002). Cognitive and psychological factors underlying secondary school students' feelings towards group work. *Educational Psychology*, 22(1), 75–91.

Carless, D. and Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325.

Chaloupskỳ, D., Chaloupská, P., and Hrušová, D. (2021). Use of fitness trackers in a blended learning model to personalize fitness running lessons. *Interactive Learning Environments*, 29(2), 213–230.

Chang, W.-L. and Yeh, Y.-c. (2021). A blended design of game-based learning for motivation, knowledge sharing and critical thinking enhancement. *Technology, Pedagogy and Education* (pp. 1–15).

Changhao, L., Majumdar, R., and Ogata, H. (2021a). Learning log-based automatic group formation: system design and classroom implementation study. *Research and Practice in Technology Enhanced Learning*, 16(1), 1–22.

Changhao, L., Toyokawa, Y., Nakanishi, T., Majumdar, R., and Ogata, H. (2021b). Supporting peer evaluation in a data-driven group learning environment. In *International Conference on Collaboration Technologies and Social Computing* (pp. 93–100). Springer.

Cheng, W. and Warren, M. (1997). Having second thoughts: Student perceptions before and after a peer assessment exercise. *Studies in Higher Education*, 22(2), 233–239.

Chowdhury, S., Endres, M., and Lanis, T. W. (2002). Preparing students for success in team work environments: The importance of building confidence. *Journal of Managerial Issues* (pp. 346–359).

Cleynen, O., Santa-Maria, G., Magdowski, M., and Thévenin, D. (2020). Peer-graded individualised student homework in a single-instructor undergraduate engineering course. *Research in Learning Technology*, 28.

Cress, U. (2008). The need for considering multilevel analysis in cscl research—an appeal for the use of more advanced statistical methods. *International Journal of Computer-Supported Collaborative Learning*, 3(1), 69–84.

Creswell, J. W., Klassen, A. C., Plano Clark, V. L., and Smith, K. C. (2011). Best practices for mixed methods research in the health sciences. *Bethesda (Maryland): National Institutes of Health*, 2013, 541–545.

Dillenbourg, P. (1999). What do you mean by collaborative learning? In *Collaborative-learning: Cognitive and Computational Approaches.* (pp. 1–19). Oxford: Elsevier.

Drury, H., Kay, J., and Losberg, W. (2003). Student satisfaction with groupwork in undergraduate computer science: do things get better? In *Proceedings of the fifth Australasian conference on Computing education-Volume 20* (pp. 77–85).

Erkens, M., Manske, S., Hoppe, H. U., and Bodemer, D. (2019). Awareness of complementary knowledge in cscl: impact on learners' knowledge exchange in small groups. In *International Conference on Collaboration and Technology* (pp. 3–16). Springer.

Fischer, F., Bruhn, J., Gräsel, C., and Mandl, H. (2002). Fostering collaborative knowledge

construction with visualization tools. *Learning and Instruction*, 12(2), 213–232.

Flanagan, B., Liang, C., Majumdar, R., and Ogata, H. (2021). Towards explainable group formation by knowledge map based genetic algorithm. In *2021 International Conference on Advanced Learning Technologies (ICALT)* (pp. 370–372). IEEE.

Forsell, J., Forslund Frykedal, K., and Hammar Chiriac, E. (2020). Group work assessment: assessing social skills at group level. *Small Group Research*, 51(1), 87–124.

Gillies, R. M. (2003). The behaviors, interactions, and perceptions of junior high school students during small-group learning. *Journal of educational Psychology*, 95(1), 137.

Gueldenzoph, L. E. and May, G. L. (2002). Collaborative peer evaluation: Best practices for group member assessments. *Business Communication Quarterly*, 65(1), 9–20.

Haq, I. U., Anwar, A., Rehman, I. U., Asif, W., Sobnath, D., Sherazi, H. H. R., and Nasralla, M. M. (2021). Dynamic group formation with intelligent tutor collaborative learning: A novel approach for next generation collaboration. *IEEE Access*, 9, 143406–143422.

Janssen, J. and Kirschner, P. A. (2020). Applying collaborative cognitive load theory to computer-supported collaborative learning: towards a research agenda. *Educational Technology Research and Development* (pp. 1–23).

Jensen, J. L. and Lawson, A. (2011). Effects of collaborative group composition and inquiry instruction on reasoning gains and achievement in undergraduate biology. *CBE—Life Sciences Education*, 10(1), 64–73.

Kanika, Chakraverty, S., Chakraborty, P., and Madan, M. (2022). Effect of different grouping arrangements on students' achievement and experience in collaborative learning environment. *Interactive Learning Environments* (pp. 1–13).

Kasch, J., van Rosmalen, P., Löhr, A., Klemke, R., Antonaci, A., and Kalz, M. (2021). Students' perceptions of the peer-feedback experience in moocs. *Distance Education*, 42(1), 145–163.

Knez, T., Dlab, M. H., and Hoic-Bozic, N. (2017). Implementation of group formation algorithms in the elars recommender system. *International Journal of Emerging Technologies in Learning (iJET)*, 12(11), 198–207.

Kyndt, E., Raes, E., Lismont, B., Timmers, F., Cascallar, E., and Dochy, F. (2013). A meta-analysis of the effects of face-to-face cooperative learning. do recent studies falsify or verify earlier findings? *Educational research review*, 10, 133–149.

Lin, H.-C., Hwang, G.-J., Chang, S.-C., and Hsu, Y.-D. (2021). Facilitating critical thinking in decision making-based professional training: An online interactive peer-review approach in a flipped learning context. *Computers & Education*, 173, 104266.

Liu, N.-F. and Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teaching in Higher education*, 11(3), 279–290.

Majumdar, R., Akçapınar, A., Akçapınar, G., Flanagan, B., and Ogata, H. (2019). Laview: Learning analytics dashboard towards evidence-based education. In *9th International Conference on Learning Analytics and Knowledge* (pp. 386–387).

Manske, S., Hecking, T., Chounta, I. A., Werneburg, S., and Ulrich Hoppe, H. (2015). Using differences to make a difference: A study on heterogeneity of learning groups. In *Computer-Supported Collaborative Learning Conference, CSCL*.

Manske, S. and Hoppe, H. U. (2016). The "Concept cloud": Supporting collaborative knowledge construction based on semantic extraction from learner-generated artefacts. In *Proceedings - IEEE 16th International Conference on Advanced Learning Technologies, ICALT 2016*.

Maqtary, N., Mohsen, A., and Bechkoum, K. (2019). Group formation techniques in computer-supported collaborative learning: A systematic literature review. *Technology, Knowledge and Learning*, 24(2), 169–190.

Mentzer, N., Laux, D., Zissimopoulos, A., and Richards, K. A. R. (2017). Peer evaluation of team member effectiveness as a formative educational intervention. *Journal of Technology Education*, 28(2), 53–82.

Meusen-Beekman, K. D., Joosten-ten Brinke, D., and Boshuizen, H. P. (2016). Effects of formative assessments to develop self-regulation among sixth grade students: Results from

a randomized controlled intervention. *Studies in Educational Evaluation*, 51, 126–136.

Mikouchi, K. A., Akita, K., and Komura, S. (2019). A critical review on project-based learning in japanese secondary education. *Bulletin of the Graduate School of Education, the University of Tokyo*, 58, 373–385.

Misiejuk, K. and Wasson, B. (2021). Backward evaluation in peer assessment: A scoping review. *Computers & Education*, 175, 104319.

Moreno, J., Ovalle, D. A., and Vicari, R. M. (2012). A genetic algorithm approach for group formation in collaborative learning considering multiple student characteristics. *Computers & Education*, 58(1), 560–569.

Motulsky, H. J. and Brown, R. E. (2006). Detecting outliers when fitting data with nonlinear regression–a new method based on robust nonlinear regression and the false discovery rate. *BMC bioinformatics*, 7(1), 1–20.

Ollesch, L., Heimbuch, S., and Bodemer, D. (2019). Towards an integrated framework of group awareness support for collaborative learning in social media. In *Proceedings of the 27th International Conference on Computers in Education* (pp. 121–130).

Ouyang, F., Chen, Z., Cheng, M., Tang, Z., and Su, C.-Y. (2021). Exploring the effect of three scaffoldings on the collaborative problem-solving processes in china's higher education. *International Journal of Educational Technology in Higher Education*, 18(1), 1–22.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. (2013). Tuned models of peer assessment in moocs. In *Proceedings of The 6th International Conference on Educational Data Mining (EDM 2013)*.

Rodríguez-Triana, M. J., Martínez-Monés, A., Asensio-Pérez, J. I., and Dimitriadis, Y. (2015). Scripting and monitoring meet each other: Aligning learning analytics and learning design to support teachers in orchestrating cscl situations. *British Journal of Educational Technology*, 46(2), 330–343.

Rohmah, K., Priyatni, E. T., and Suwignyo, H. (2021). Assessment of learning development to improve student's appreciative and critical thinking abilities in drama appreciation course. In *4th Sriwijaya University Learning and Education International Conference (SULE-IC 2020)* (pp. 495–502). Atlantis Press.

Sadeghi, H. and Kardan, A. A. (2016). Toward effective group formation in computer-supported collaborative learning. *Interactive learning environments*, 24(3), 382–395.

Salihoun, M., Guerouate, F., Berbiche, N., and Sbihi, M. (2017). How to assist tutors to rebuild groups within an its by exploiting traces. case of a closed forum. *International Journal of Emerging Technologies in Learning*, 12(3).

Sánchez, O. R., Ordóñez, C. A. C., Duque, M. Á. R., and Pinto, I. I. B. S. (2021). Homogeneous group formation in collaborative learning scenarios: An approach based on personality traits and genetic algorithms. *IEEE Transactions on Learning Technologies*, 14(4), 486–499.

Sanz-Martínez, L., Er, E., Martínez-Monés, A., Dimitriadis, Y., and Bote-Lorenzo, M. L. (2019). Creating collaborative groups in a mooc: a homogeneous engagement grouping approach. *Behaviour & Information Technology*, 38(11), 1107–1121.

Šerić, M. and Garbin Praničević, D. (2018). Managing group work in the classroom: An international study on perceived benefits and risks based on students' cultural background and gender. *Management: Journal of Contemporary Management Issues*, 23(1), 139–156.

Shiho, N. (2021). A study on subjectivity and interactive dialogue in lessons (i): Critical examination of "proactive, interactive and authentic learning". *Bulletin of the Graduate School of Education and Human Development (Educational Sciences) Nagoya University*, 68(1), 25–37.

Siemens, G. (2012). Learning analytics: envisioning a research discipline and a domain of practice. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 4–8).

Stahl, G., Koschmann, T., and Suthers, D. D. (2006). Computer-supported collaborative learning: An historical perspective. *Cambridge handbook of the learning sciences* (pp. 409–426).

Strauß, S. and Rummel, N. (2021). Promoting regulation of equal participation in online

collaboration by combining a group awareness tool and adaptive prompts. but does it even matter? *International Journal of Computer-Supported Collaborative Learning*, 16(1), 67–104.

Strijbos, J.-W. (2010). Assessment of (computer-supported) collaborative learning. *IEEE transactions on learning technologies*, 4(1), 59–73.

Sukstrienwong, A. (2017). A genetic-algorithm approach for balancing learning styles and academic attributes in heterogeneous grouping of students. *International Journal of Emerging Technologies in Learning*, 12(3).

Tharim, A. H. A., Mohd, T., Othman, N. A., Nasrudin, N. H., Jaffar, N., Shuib, M. N., Kurdi, M. K., Yusof, I., et al. (2016). Peer evaluation system in team work skills assessment. In *7th International Conference on University Learning and Teaching (InCULT 2014) Proceedings* (pp. 603–616). Springer.

Toyokawa, Y., Majumdar, R., Lecailliez, L., Liang, C., and Ogata, H. (2021). Technology enhanced jigsaw activity design for active reading in english. In *2021 International Conference on Advanced Learning Technologies (ICALT)* (pp. 367–369). IEEE.

Urhahne, D., Schanze, S., Bell, T., Mansfield, A., and Holmes, J. (2010). Role of the teacher in computer-supported collaborative inquiry learning. *International Journal of Science Education*, 32(2), 221–243.

van der Velde, M., Sense, F., Borst, J., and van Rijn, H. (2021). Alleviating the cold start problem in adaptive learning using data-driven difficulty estimates. *Computational Brain & Behavior*, 4(2), 231–249.

Van Leeuwen, A. (2015). Learning analytics to support teachers during synchronous cscl: Balancing between overview and overload. *Journal of learning Analytics*, 2(2), 138–162.

Van Leeuwen, A., Janssen, J., Erkens, G., and Brekelmans, M. (2014). Supporting teachers in guiding collaborating students: Effects of learning analytics in cscl. *Computers & Education*, 79, 28–39.

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard university press.

Wang, M., Guo, W., Le, H., and Qiao, B. (2020). Reply to which post? an analysis of peer reviews in a high school spoc. *Interactive Learning Environments*, 28(5), 574–585.

Wang, Q. (2010). Using online shared workspaces to support group collaborative learning. *Computers & Education*, 55(3), 1270–1276.

Wessner, M. and Pfister, H.-R. (2001). Group formation in computer-supported collaborative learning. In *Proceedings of the 2001 international ACM SIGGROUP conference on supporting group work* (pp. 24–31).

Willey, K. and Gardner, A. (2010). Investigating the capacity of self and peer assessment activities to engage students and promote learning. *European Journal of Engineering Education*, 35(4), 429–443.

Xethakis, L. J. (2018). Psychometric adaptation of a japanese version of the feelings towards group work questionnaire for use in the japanese sla context. *Kumamoto University studies in social and cultural sciences*, 16, 219–247.

Yoshida, M., Xiong, C., Liu, Y., and Liu, H. (2020). An investigation into the formation of learning groups on social media and their growth. *Interactive Learning Environments* (pp. 1–14).

Zamani, M. (2016). Cooperative learning: Homogeneous and heterogeneous grouping of iranian efl learners in a writing context. *Cogent Education*, 3(1), 1149959.

Zhang, L., Ayres, P., and Chan, K. (2011). Examining different types of collaborative learning in a complex computer-based environment: A cognitive load approach. *Computers in Human Behavior*, 27(1), 94–98.

Appendix A. Details of Mann-Whitney U tests

Table A1. Peer ratings of groups formed by heterogeneous algorithm and random arrangement

	Group composition	N	Mean	SD	p	effect size
Subjectivity	Heterogeneous	39	3.876	0.976	0.520	0.087
	Random	35	3.706	1.118		
Communication	Heterogeneous	39	3.769	1.012	0.445	0.103
	Random	35	3.535	1.224		
Learning	Heterogeneous	39	3.829	0.983	0.051	0.263
	Random	35	3.368	1.165		

Table A2. Self-perception of group learning survey of groups formed by heterogeneous algorithm and random arrangement

	Group composition	N	Mean	SD	p	effect size
Self-perception	Heterogeneous	34	18.206	4.176	0.831	-0.031
	Random	32	18.176	4.421		

Table A3. Peer ratings of groups formed by homogeneous algorithm and random arrangement

	Group composition	N	Mean	SD	p	effect size
Subjectivity	Homogeneous	72	4.171	0.720	0.003**	0.291
	Random	70	3.713	1.024		
Communication	Homogeneous	72	4.030	0.807	0.037*	0.202
	Random	70	3.658	1.074		
Learning	Homogeneous	72	4.191	0.707	<0.001***	0.354
	Random	70	3.677	0.963		

* $p < .05$, ** $p < .01$, *** $p < .001$.

Table A4. Self-perception of group learning survey of groups formed by homogeneous algorithm and random arrangement

	Group composition	N	Mean	SD	p	effect size
Self-perception	Homogeneous	68	19.324	2.878	0.003**	0.305
	Random	58	17.483	3.521		

** $p < .01$.

Table A5. Peer ratings of groups created by homogeneous and heterogeneous algorithms in comparative reading context

	Group composition	N	Mean	SD	p	effect size
Subjectivity	Heterogeneous	70	3.735	0.862	0.662	-0.051
	Homogeneous	37	3.712	1.022		
Communication	Heterogeneous	70	3.689	0.934	0.293	-0.124
	Homogeneous	37	3.716	1.190		
Learning	Heterogeneous	70	3.783	0.731	0.635	-0.056
	Homogeneous	37	3.676	1.155		

Table A6. Self-perception of groups created by homogeneous and heterogeneous algorithms in comparative reading context

	Group composition	N	Mean	SD	p	effect size
Self-perception	Heterogeneous	50	18.420	3.818	0.297	0.141
	Homogeneous	29	17.138	5.370		

Table A7. Peer ratings of groups created by homogeneous and heterogeneous algorithms in idea exchange context

	Group composition	N	Mean	SD	p	effect size
Subjectivity	Heterogeneous	39	3.876	0.976	0.119	-0.178
	Homogeneous	72	4.171	0.720		
Communication	Heterogeneous	39	3.769	1.012	0.097	-0.190
	Homogeneous	72	4.030	0.807		
Learning	Heterogeneous	39	3.829	0.983	0.042*	-0.233
	Homogeneous	72	4.191	0.707		

*p < .05.

Table A8. Self-perception of groups created by homogeneous and heterogeneous algorithms in idea exchange context

	Group composition	N	Mean	SD	p	effect size
Self-perception	Heterogeneous	34	18.206	4.176	0.108	-0.194
	Homogeneous	68	19.324	2.878		