

Doctoral Dissertation

GLOBE: Data-Driven Support for Group Learning

LIANG CHANGHAO

Supervisor: Professor Hiroaki OGATA

August 24, 2023

Department of Social Informatics
Graduate School of Informatics
Kyoto University

Abstract

Collaborative learning has become increasingly popular in various educational contexts, benefiting the development of soft skills such as critical thinking, problem-solving, and interpersonal communication, which are highly valued in modern society. Typically, collaborative learning occurs in the form of small-group activities. With the introduction of tablets in Japanese classrooms, computer-supported collaborative learning (CSCL) and learning analytics (LA) provide digital tools and data support, creating immense opportunities to enhance these activities with information technologies.

However, obstacles to providing a valid scaffold for group learning still exist. 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. Students of traditional classrooms seldom use digital tools for group learning, which leads to a cold start problem for the lack of enough learning logs to create learner models that can be used to allocate students based on their attributes. Even with the support of computers, there remains a chance that teachers would get overwhelmed if they are not familiar with the computer-supported tools for orchestration. In addition, to evaluate the performance of the group work, only the teacher's evaluation is not enough since one teacher cannot check what is happening for all groups during group learning. Currently, many researchers focus on utilizing LA tools during the orchestration phase of the group work, while valid support for group formation and evaluation phases deserves further attention.

To address these issues, this research proposes the Group Learning Orchestration Based on Evidence (GLOBE) framework, which supports group learning in various contexts using data-driven systems. The aim is to apply LA to the CSCL process, consolidating various learning log data to support each phase of group activities and figure out predictors of successful group work from these inputs. The research introduces two key innovations: the utilization of multiple data sources for group formation through genetic algorithms, and the implementation of learning analytics for optimized parameter selection and purpose-based recommendations to teachers by employing evidence of continuous multiple group learning activities.

The implementation of the GLOBE framework takes three steps: synthesizing data, utilizing data, and analyzing data. Firstly, a group formation system using genetic algorithms is designed and implemented to form groups using learning log data from various

sources. Annotation data of common markers can also be reflected in group formation. Secondly, a continuous data-driven support paradigm for the entire group learning process is proposed, incorporating input from peer and teacher evaluations for subsequent groupings. Although only a few student model data were used in the current studies, further opportunities for LA-enhanced group work orchestration were revealed following the continuous data flow, even in classroom-based contexts where no initial data existed. Furthermore, by utilizing accumulated group learning evidence in the GLOBE ecosystem, predictive group formation indicators were explored that can enable automatic group formation based on teachers' objectives in different contexts for desirable performance in subsequent group learning activities.

Empirical studies show that the GLOBE framework provides a low threshold for teachers to adapt the data-driven workflow for group learning design, promoting the use of digital systems in routine practice. Implementations of GLOBE systems have shown that they can reduce the time for teachers and students from trivial works of group formation and evaluation. Additionally, a new perspective is suggested to explore how group composition with diverse student model data tends to make a difference in the performance of subsequent phases. By investigating specific student model variables for group formation, heterogeneity can be inspected among which characteristics weigh more to affect the subsequent group learning process and outcome. This doctoral dissertation introduces the group formation and peer evaluation modules in the GLOBE framework and the continuous data flow, with four empirical studies corresponding to the three data-driven steps of GLOBE implementation.

Contents

1	Introduction	1
1.1	Enhancing collaborative learning with CSCL and Learning Analytics . . .	1
1.2	Current issues on group learning practice	2
1.3	Proposed solution	3
2	Literature Review	7
2.1	Group learning attributes and indicators	7
2.2	Group formation	9
2.3	Group work evaluation	12
3	Group learning based on evidence (GLOBE) framework and systems	14
3.1	Group formation: algorithmic grouping using logs in student model	14
3.2	Group work evaluation: peer evaluation and feedback	20
3.3	Supporting continuous data-driven group works under GLOBE	20
4	Synthesize: Group formation using learning log data	24
4.1	Study 1: Group formation based on knowledge and relationship	25
4.1.1	Aim and research question	25
4.1.2	Learning Context and participants	25
4.1.3	Research Design	26
4.1.4	Procedure	26
4.1.5	System Usage	28
4.1.6	Data Collection	28
4.1.7	Data Analysis	29
4.1.8	Result and Inferences	29
4.1.9	Discussion	33

4.1.10	Conclusion and Future Work	38
4.2	Study 2: Group formation using reading annotations	40
4.2.1	Aim and research questions	40
4.2.2	System innovation: Reading marker attributes for group formation	42
4.2.3	Study Context and participants	43
4.2.4	Research design	44
4.2.5	Procedure	44
4.2.6	Data collection	46
4.2.7	Data analysis	47
4.2.8	Composition of marker-based groups	48
4.2.9	Results	49
4.2.10	Discussion	51
4.2.11	Conclusion and future work	57
5	Utilize: Using group evaluation for subsequent group work	59
5.1	Study 3: Group formation using continuously accumulated peer rating data	59
5.1.1	Aim and research question	59
5.1.2	Study context and design	60
5.1.3	Participants	62
5.1.4	Procedure	62
5.1.5	Instruments and data collection	63
5.1.6	Data analysis	64
5.1.7	Results	65
5.1.8	Discussion	71
5.1.9	Conclusion and Future work	77
6	Analyze: Recommendation of optimal group formation settings	79
6.1	Study 4: Predictive group work indicators for optimal group formation settings	79
6.1.1	Aim and research questions	79
6.1.2	Research context and participants	80
6.1.3	Procedure	80
6.1.4	Data collection	81

6.1.5	Data analysis	81
6.1.6	Results	83
6.1.7	Discussion	85
6.1.8	Conclusion and future work	90
7	Discussion	91
7.1	Summary of research	91
7.2	Implications	93
7.3	Limitations	94
7.4	Future work	95
8	Conclusion	96
	Acknowledgement	97
	References	99

List of Figures

1.1	Thesis overview: associated problems, solutions and academic publications.	6
2.1	Collaborative Process Attributes and example indicators (Janssen & Kirschner, 2020)	7
3.1	GLOBE framework and its implementation systems	15
3.2	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)	16
3.3	The mapping of student model variable values to the student characteristic representation matrix (Flanagan et al., 2021)	17
3.4	Algorithms and parameters used in the group formation module	17
3.5	Creating and visualizing friendship for group formation.	18
3.6	Example of the group formation details of a heterogeneous group and a homogeneous group (raw scores scaled to 0-100)	19
3.7	Interface of peer rating with three criteria set by the teacher	21
3.8	Interface of peer feedback panel	22
3.9	Interface of peer rating with three criteria set by the teacher	22
3.10	Example of a continuous data-driven support data flow under GLOBE	23
3.11	Example of a visualization of the weighted score considering the reliability of peer ratings	23
4.1	Process of the in-class group work.	27
4.2	Transition patterns of utterance duration in knowledge exchange phase between activity A3 and A5.	31
4.3	Transition patterns of utterance duration for knowledge exploration phase activity between activity A3 and A5.	33

4.4	Difference in affective state scores for knowledge exchange phase.	34
4.5	Typical workflow for activities involving regrouping.	36
4.6	Typical workflow for flipped reading activities.	37
4.7	Student attributes from assorted resources for algorithmic group formation (Liang, Majumdar, & Ogata, 2021)	41
4.8	Workflow to feature marker attributes for group formation	42
4.9	Procedure of the study	45
4.10	Rubrics used by teacher to grade summary	46
4.11	Box plot comparing learning gains in the vocabulary quizzes under three conditions	50
4.12	Transition graph of vocabulary quiz scores	52
4.13	Example procedures for group work implementation under GLOBE frame- work and LEAF.	57
5.1	Idea exchange group learning: Classroom implementation workflow	60
5.2	Procedure of the group learning experiment	61
5.3	Box plot comparing heterogeneity of groups created by three approaches	66
6.1	Workflow of the weekly activity implemented in the course	82
6.2	Results of correlation analysis of individual-level indicators of group work	84
6.3	Results of correlation analysis of group-level indicators of group work	86
6.4	Suggested group formation strategies based on correlations between group- level attributes	88
6.5	System innovation: From parameterized grouping to automatic grouping	89
7.1	Summary of research from the data-driven perspective.	92
7.2	Summary of research from the data-driven perspective.	92

List of Tables

2.1	Group formation in Computer Supported Cooperative Work (CSCW). . . .	10
2.2	Compilation of notable studies on group formation in CSCL.	11
4.1	Summary of data collection.	26
4.2	Difference in engagement indicators for knowledge exchange phase.	30
4.3	Cutoff of three strata of the utterance duration transition graph.	30
4.4	Difference in engagement indicators for knowledge exploration phase activity.	32
4.5	ANOVA test for the scores of the lastest exam of three classes	44
4.6	Details of marker-based heterogeneous groups	48
4.7	Kruskal-Wallis test of the performance scores of the summary assignment .	49
4.8	Dunn’s Post Hoc Comparisons - scores of the summary assignment	49
4.9	Survey items and Kruskal-Wallis test of responses	53
5.1	5-item survey on the self-perception of group work (adapted from Drury et al. (2003))	64
5.2	ANOVA of pre-test score and attitude towards group learning survey . . .	65
5.3	Descriptive statistics and ANOVA of group heterogeneity under three group formation approaches	66
5.4	Post Hoc Comparisons of groups formed by different approaches	67
5.5	Overall results of comparative studies of groups created by data-driven algorithmic group formation and random arrangement	68
5.6	Peer ratings of groups formed by the heterogeneous algorithm and random arrangement	68
5.7	Self-perception of group learning survey of groups formed by the heteroge- neous algorithm and random arrangement	69

5.8	Peer ratings of groups formed by the homogeneous algorithm and random arrangement	69
5.9	Self-perception of group learning survey of groups formed by the homogeneous algorithm and random arrangement	69
5.10	Overall results of comparative studies of groups created by heterogeneous and homogeneous algorithms	70
5.11	Peer ratings of groups created by homogeneous and heterogeneous algorithms in idea exchange context	70
5.12	Self-perception of groups created by homogeneous and heterogeneous algorithms in idea exchange context	71
5.13	Peer ratings of groups created by homogeneous and heterogeneous algorithms in comparative reading context	71
5.14	Self-perception of groups created by homogeneous and heterogeneous algorithms in comparative reading context	71
6.1	Group formation and group work topics in the course	81
6.2	Indicators used in this study	83

Chapter 1

Introduction

1.1 Enhancing collaborative learning with CSCL and Learning Analytics

Collaborative learning has become increasingly popular in various educational settings as it benefits the development of soft skills, such as critical thinking, problem-solving, and interpersonal communication, that are highly valued in modern society (Dinh et al., 2021; Stahl et al., 2006). During collaborative learning, participants work together to share ideas, help each other or accomplish team goals (Dillenbourg, 1999). Typically, this type of learning occurs in the form of small-group activities (Gillies, 2016). With the introduction of tablets in Japanese classrooms, there has been accelerated progress in promoting the integration of educational ICT environments in routine practice. The ongoing GIGA school project in Japan and the impact of the COVID-19 pandemic, have expedited progress in the advancement and implementation of educational ICT environments. As a result, computer-supported collaborative learning (CSCL) (Stahl et al., 2006) and learning analytics (LA) (Siemens, 2012) can provide digital tools and data support, creating immense opportunities to enhance these activities with information technologies.

Computer-Supported Collaborative Learning (CSCL) is an emerging branch of learning sciences concerned with studying how people learn together with the help of computers (Stahl et al., 2006). It is founded on the premise that technology can effectively assist in the process of collaborative knowledge construction and problem solving (Jeong et al., 2019). The application of CSCL runs through a broad variety of contexts throughout the process from creating groups, group regulation, and in-group interaction to group evaluation and reflection. For instance, kit-map generation is a typical activity where

CSCL is frequently utilized for brainstorming and knowledge building (Manske & Hoppe, 2016). Workshop such as programming projects is another application (Moreno et al., 2012) where students harvest collaboration skills.

Meanwhile, in recent years, there has been a broad implementation of computerized teaching, smart tutoring systems, and artificial intelligence methodologies, leading to the generation of abundant data related to student learning behavior (Picciano, 2012). Accordingly, learning analytics (LA) has been introduced to measure, collect, analyze, and report data about learners and their contexts, with the aim of improving their learning environment (Siemens, 2012). By utilizing previous student-produced learning log data, it is possible to conduct predictive analytics in assorted educational settings (Chen et al., 2021; Ferguson, 2012), which provides an opportunity to scaffold CSCL as well. Such analysis can affect their learning behaviors and improve outcomes by proper remedial actions (Banihashem et al., 2022; Ifenthaler & Yau, 2020).

1.2 Current issues on group learning practice

Despite the increasing popularity of group learning in pedagogical practices, there are still gaps in the use of technical support, particularly during the group formation and evaluation phases. To orchestrate a successful group learning activity, teachers must envision the lesson, enable collaboration, encourage students, ensure learning, and evaluate achievements. However, this process can be time-consuming. During the group formation phase, teachers need to compose each group and align students according to various learning contexts (Urhahne et al., 2010). They often spend significant time on trivial group formation work and may struggle to create appropriate groups in contexts such as MOOCs (C. Wang & Xu, 2023). Appropriate group formation can be challenging for teachers hence they may resort to random grouping or pairing neighboring students due to the difficulties involved in forming groups in real-time (Salihoun et al., 2017).

Furthermore, even in a CSCL-supported context such as the LA-enhanced group formation system, teachers still need to understand and select parameters. This may be confusing for them due to their unfamiliarity with digital systems. These technology adoption barriers may require additional effort from teachers and distract them from initiating classroom activities (Austin et al., 2010). To simplify this process for teachers and ease the burden of complex parameter selections, it is recommended to automatically

recommend appropriate group formation indicators that are predictive of desirable group work performance (Slof et al., 2021).

Evaluating the performance of in-class group work can be also challenging, as teachers may not have real-time support to monitor all groups simultaneously, leading to difficulties in providing fair evaluations for all students (Amarasinghe et al., 2021; Kasch et al., 2021). Teachers may also face issues related to social loafing and free-riding, which further complicate the evaluation process (Q. Wang, 2010). Self-assessment and peer-assessment methods are often adopted as alternatives (Forsell et al., 2020). Currently, many researchers focus on the implementation of LA tools during the ongoing 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.

In addition, there is a "cold start" problem due to the lack of sufficient learning logs for group allocation, especially in traditional classroom contexts where students rarely use digital tools (Brusilovsky et al., 2015; Pliakos et al., 2019). While many researchers have focused on the topic of group formation, existing studies have mainly examined the effects of specific indicators and algorithms based on their group learning objectives. These studies often collect one-time data to create groups using fixed characteristics and techniques in a controlled experimental environment, with an emphasis on revealing the causal relationship of the intervention, i.e., internal validity (Kuromiya et al., 2020). Few studies have established a sustainable environment that encompasses student model variables in multiple learning platforms for group formation tasks in various contexts, where data comes from routine practice settings (Maissenhaelter et al., 2018). In other words, related studies on group work support tend to focus on specific experimental settings but rarely consider the data-driven perspective.

1.3 Proposed solution

The advancement of information infrastructures, coupled with increasing learning log data, offers solutions to the obstacles faced in the group learning process. The emergence of integrated online learning platforms has facilitated reading activities and collaborative learning through the provision of a data-driven environment for recording, analyzing, and visualizing learners' actions (Ogata et al., 2022). This learning environment offers

immense opportunities to support various learning activities. In a data-driven platform, learning logs from various sub-systems can be connected and aggregated in a comprehensive repository, allowing all sub-services of the platform to utilize them (Kuromiya et al., 2020). These real-world data hold the advantage of generalizability (Maissenhaelter et al., 2018) and convenience for future extraction of evidence (Kuromiya et al., 2020).

The Group Learning Orchestration Based on Evidence (GLOBE) infrastructure is pro-posed in this research to support group work with data-driven systems. It integrates digital systems to create a data-driven environment based on the LEAF (Learning Analytics Framework) (Ogata et al., 2023). The thesis focuses on the design of the data-driven system and its empirical implementations surrounding the phases of GLOBE with the iterative data flow. Several empirical studies were conducted to investigate the impact of GLOBE systems in different learning contexts and figure out predictors of successful group work from these inputs to orchestrate an ecosystem, providing a versatile foundation that can adapt to various contexts and effectively support group learning.

This research introduces several innovative ideas and findings. Firstly, it explores the use of multiple data sources such as daily e-book reading behaviors in group formation using Genetic Algorithms (GA), addressing the limitations of traditional group formation based solely on scores. The proposed system also incorporates relationship data and reading marker overlap data, enabling the consideration of mutual relationship indicators within groups. These additions make group compositions explainable based on both individual attributes and relationship indicators. The findings indicate that the system is capable of immediately suggesting unexpected group combinations which have proven to be successful in empirical studies, thereby reducing teachers' bias on group formation. Secondly, the study proposes learning analytics for optimized parameter selection and purpose-based recommendations for teachers by employing evidence of continuous multiple group learning activities. Some new combinations of group formation settings are identified in academic reading contexts, which align with existing theories and contribute to the understanding of effective group formation strategies and their impact on learning outcomes.

The thesis is organized from a data perspective, where the implementation of the GLOBE framework takes three steps: synthesizing data, utilizing data, and analyzing data. Firstly, a group formation system using genetic algorithms is designed and implemented to form groups using learning log data from various sources. Secondly, a con-

tinuous data-driven support paradigm for the entire group learning process is proposed, incorporating input from peer and teacher evaluations for subsequent groupings. Further, with the accumulation of group learning data, predictive group formation indicators were explored to enable automatic group formation based on teachers' objectives in different contexts for desirable performance in subsequent group learning activities. The main research questions are formulated as follows, and an overview of the thesis is presented in Figure 1.1.

Topic 1: How to design an algorithmic group formation tool using multiple learner model attributes?

Topic 2: How to support the whole group learning process with continuous data workflow?

Topic 3: How to achieve automatic group formation and predict desirable group work in subsequent phases?

The subsequent chapters are organized as follows: Chapter 2 introduces the theoretical and technical infrastructures from related works. Based on this, Chapter 3 provides a detailed description of the GLOBE framework and its composition systems. Chapters 4 to 6 present the empirical studies conducted in different learning contexts, following the three steps of data leverage. The learning scenarios vary from primary schools, middle schools to higher education levels. In Chapter 7, a general discussion of the implications and limitations of this research is provided. Finally, in Chapter 8, the thesis concludes by summarizing the findings of our research.

Associated Problems	Teachers usually take much time on trivial group formation work and can be unfamiliar with the students. Traditional grouping strategies have shortcomings.	Could start problem for the lack of enough learning logs for group allocation. Related studies on data-driven group work addressed single episode, but rarely focus on the continuous lifecycle.	Teachers can get overwhelmed when using digital systems Even with LA-enhanced group formation system, teacher still need to understand and select parameters.
Research Topics	Topic 1: How to design an algorithmic group formation tool using multiple learner model attributes?	Topic 2: How to support the whole group learning process with continuous data workflow?	Topic 3: How to achieve automatic group formation and predict desirable group work in subsequent phases?
Solutions	Study 1 - Group formation system using teacher grading scores and relationship Study 2 - Group formation based on reading annotations of content	Study 3 - Group formation using continuously accumulated peer rating data	Study 4 - Predictive group work indicators for optimal group formation settings
Chapters	4 - Synthesize: Group formation using learning log data	5 - Utilize: Using group evaluation for subsequent group work	6 - Analyze: Recommendation of optimal group formation settings
Academic Publications	Study 1: RPTTEL (2021) Study 2: JLCE (2023)	Study 3: ILE (2022)	Study 4: ET&S (2023)

Figure 1.1.: Thesis overview: associated problems, solutions and academic publications.

Chapter 2

Literature Review

2.1 Group learning attributes and indicators

When conducting group learning in pedagogy contexts, multiple issues should be considered in different stages (Urhahne et al., 2010). To characterize these issues with data, multiple indicators are proposed which reveal certain aspects of group work. Janssen and Kirschner (2020) put forward the concept of Collaborative Process Attributes that depict collaboration in three constructs: antecedents, processes, and consequences (see Figure 2.1). Indicators of antecedent attributes can pose an effect on processes and consequences of collaboration. However, which antecedent attributes influence the process and consequences of collaboration more was less discussed in previous studies, though it can be not only instructive for system innovation on automatic grouping but also assist teachers in setting groups appropriately with assorted student model data. In a digital learning environment with abundant learning log data, many of these indicators are recorded as learner models that depict the learning characteristics of students (Brusilovsky et al., 2015).

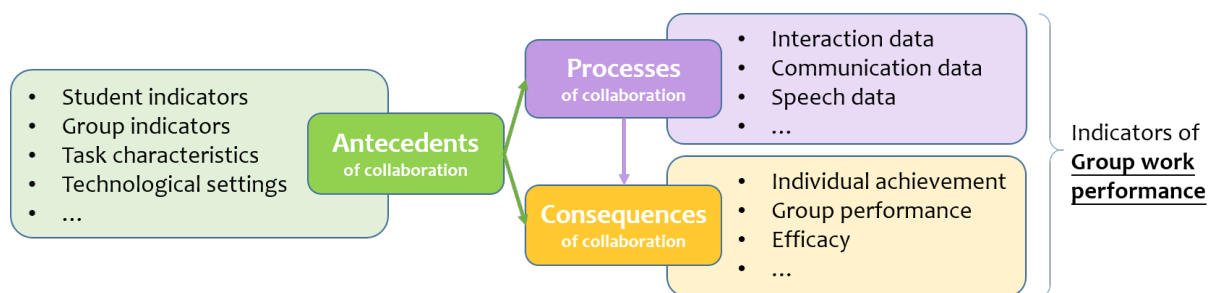


Figure 2.1: Collaborative Process Attributes and example indicators (Janssen & Kirschner, 2020)

For antecedents of collaboration, Janssen and Kirschner (2020) presented several typical instances based on what it describes. Student and group indicators are frequently-discussed (Saqr et al., 2020) and are prone to vary from group work tasks. Student indicators encompass all domain-specific and domain-independent information and as quantified indicators (Boticki et al., 2019), which can be easily derived from student model attributes under data-driven infrastructures. For example, gender, previous knowledge and task experience, preferences of learning styles, and personalities can be enveloped in the student indicators of group work (Abnar et al., 2012; Sánchez et al., 2021; Savicki et al., 1996; Zheng & Pinkwart, 2014). Group indicators describe characteristics of groups such as group size and intimacy (Amason & Sapienza, 1997; Huckman et al., 2009). Meanwhile, the heterogeneity distribution of student indicators within one group was also highlighted (Xu et al., 2020), which is closely connected to data-driven algorithmic group formation.

Processes of collaboration are an essential part of CSCL research (Strode et al., 2022) since they can offer a holistic picture of the collaborative process that records the evidence during group work. The communication data, no matter in the form of oral utterance (Donnelly et al., 2017) or online forums (Fidalgo-Blanco et al., 2015), assumes widely-used group learning evidence in related studies. Timeline sequence modeling, social network analysis (SNA), and epistemic network analysis (ENA) are conducted to further investigate the interaction data (Fidalgo-Blanco et al., 2015; Hoppe et al., 2021; Kanika et al., 2022). Using these interaction data during group work, it is feasible to use machine learning techniques to predict group performance (Cen et al., 2016). However, these data get available only when the current group work has started and the groups have been created.

Consequences of collaboration disclose the outcome of collaborative learning (Janssen & Kirschner, 2020). On the one hand, individual achievement estimates how much one has learned throughout the group work, especially for cognitive skills and knowledge acquisition. On the other hand, group performance is another indicator of collaboration quality, which can include the scores of group presentations and collaboratively composed reports.

Related research on data-supported group learning investigated the impact of specific student and group indicators in controlled experiments. For instance, previous knowledge and task experience proved to be closely related to group work performance in a collaborative programming context (Hsu et al., 2021; Rentsch & Klimoski, 2001). Similarly, Xu et al. (2020) also found the education level and domain knowledge of users can

interactively predict users' knowledge gained in collaborative web searching sessions. In parallel, the heterogeneity of a group also affects the group work performance (Sánchez et al., 2021), and the impact of group heterogeneity can be different depending on the learning context (Manske et al., 2015).

However, existing studies have mostly examined specific indicators of group work performance in isolated studies, which only described a single episode and had limited insight into their comprehensive evaluation and relative significance. Few investigations have taken a holistic approach to examine the underlying attributes comprehensively and ascertain their respective weights. Further, the potential for a continuous life cycle that enables the reuse of group learning data for ongoing improvement in multiple rounds has received less attention in the literature.

2.2 Group formation

In a broader view of Computer Supported Cooperative Work (CSCW), researches on group formation focus on "studying and designing technologies that bring people together in partnerships, teams, crowds, communities, and other collectives" (Harris et al., 2019). The applications of group formation span across various domains, as listed in Table 2.1. These applications include team building, expert locating, and partner matching, among others, and are relevant in areas such as expert identification within enterprises, academic collaborations, and multiplayer video games.

In the learning and education field, group formation is a fundamental task to set about a group learning task (Wessner & Pfister, 2001). Collaborative learning with properly formed groups outperforms traditional teaching (Kyndt et al., 2013), while improperly used group formation parameters may raise several problems that lead to failure (Q. Wang, 2010). Various issues such as group members' characteristics, the context of the grouping process and the techniques used to form the group could affect the group learning processes (Maqtary et al., 2019).

The characteristics of students lay the foundation to perform group formation algorithms. These student characteristics correspond to the antecedent attributes in the previous section and can be acquired in online learning platforms where multiple learning log data are accumulated. In the data-rich environment, student model data makes it possible to take student characteristics into account when creating groups (Boticki et al.,

Table 2.1: Group formation in Computer Supported Cooperative Work (CSCW).

Field	Delivery	User attributes
Enterprise (McDonald & Ackerman, 2000)	Expertise locating system to find experts inside a company	Explicit Ratings, user behaviors, etc.
Academic collaboration (Heck, 2013)	Recommendation system to facilitate the process of identifying and finding the right colleagues	Social information gleaned from citations and reference data
Video games (Benefield et al., 2016; Y. J. Kim et al., 2017)	Analysis to predict Team effectiveness and performance	In-game social networks, collective intelligence
Education	Group formation for Collaborative learning	Learner model attributes (Brusilovsky et al., 2015)

2019).

The context is important as well since the optimal settings of group formation can differ from the purpose and traits of group work activity. For example, learning with peer help calls for heterogeneity of knowledge level according to the ZPD theory from Vygotsky (1980), while homogeneous groups perform better in situations that encourage interaction and familiarity of group mates (Salihoun et al., 2017; Sanz-Martínez et al., 2019).

Manifold techniques were employed for learning group creation based on different student model data and purposes. Table 2.2 presents a compilation of notable studies on group formation in CSCL. One approach for creating groups with unbalanced abilities is to rank students according to a specific indicator and select students from different parts of the distribution (Haq et al., 2021). Clustering techniques underpinned by distance measurements are used for homogeneous groupings, such as the K-means algorithm that puts students in the same cluster together (Amara et al., 2016; Manske et al., 2015) and hierarchical clustering for group recommendations (Chang et al., 2017). In cases where students created abundant learner-generated content, the semantic method can group students (Isotani et al., 2009) based on textual features in terms of knowledge diversity, textual similarity as well as a semantic network of learner’s interaction texts (Erkens et al., 2019; Manske & Hoppe, 2016). However, expressing the heterogeneity of groups in

comparable values proves challenging when using semantic matchmakers (Konert et al., 2014).

Table 2.2: Compilation of notable studies on group formation in CSCL.

Technique	Context	Student attributes
Dynamic grouping based on score ranks (Haq et al., 2021)	Online classes of object-oriented programming	Individual quiz scores and learning style survey
Clustering and distance method (Manske et al., 2015)	Inquiry-based learning scenario with concept map and text writing	Artifacts, particularly learning objects and the assessment of motivational scores
Semantic method (Erkens et al., 2019; Manske & Hoppe, 2016)	Collaborative knowledge map generation	Learner-generated artifacts, texts, concept maps and hypotheses
Multi-objective optimized genetic method (Moreno et al., 2012)	Course subject discussion of a computer programming	Student knowledge levels, student communicative skills, student leadership skills

To deal with group formation from multiple student attributes, Moreno et al. (2012) put forward a genetic algorithm (GA) that can generate different group compositions (heterogeneous or homogeneous) in light of the calculated fitness values. One merit of the genetic algorithm is its flexibility in the number and type of attributes. The fitness values can be estimated by distance measures of vectors such as the sum of the squared differences, which can reflect the heterogeneity of the student characteristics. In this way, homogeneous groups consisting of similar group members, or heterogeneous groups with dissimilar group members can be determined. The genetic algorithm presents flexibility owing to the fitness functions that can be adjusted to meet various grouping purposes and accommodate assorted input variables as was discussed in Flanagan et al. (2021) and Revelo Sánchez et al. (2021). Krouska et al. (2023) went further on the crossover in the iteration process.

To develop a group formation system with intelligent algorithms, Konert et al. (2014) pointed out four criteria that guide the system design: (1) flexible parameter selection depending on contexts, (2) availability of several algorithms (homogeneous, heterogeneous, and mixed), (3) assessment and optimization of group formation, (4) minimization of differences among groups. Based on this guidance, GroupAL was put forward using a

similar technique of vector optimization as GA. The GroupAL algorithm also provides flexible settings of parameters and criteria (heterogeneous or homogeneous) to meet different learning scenarios. Similar to the fitness function in GA, the optimal group allocation also relies on the defined metrics that depict the distance among participants and pairwise disjoint groups. However, without multiple iterations implemented in GA, GroupAL assigns participants to learning groups only once. Under the same criteria and parameter settings, both GroupAL and GA can make different cohorts of groups since both approaches start from a randomized group allocation. Further, there were efforts of data integration to derive data from e-learning systems such as MoodlePeers as extensions of the GroupAL project (Konert et al., 2016). These endeavors highlight the potential of integrating the group formation system into existing platforms with abundant learning log data, enabling further research on optimal group formation settings.

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 (Strauß & Rummel, 2021), and enriched group awareness information (Ollesch et al., 2019) were adopted to support the group work process.

Nevertheless, 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; Van Leeuwen, 2015). Meanwhile, problems of social loafing and free riding (Strijbos, 2010) are prevalent that remain large obstacles to successful group learning activities. Therefore, peer evaluation becomes imperative to alleviate teachers' workload and provide a real-time inspection across the group learning process (Willey & Gardner, 2010). The peer evaluation tools evolve from paper-based surveys to digital files and online platforms (Cleynen et al., 2020; Tharim et al., 2016), making the evaluation delivery process faster Cleynen et al. (2020) with anonymity (Cheng & Warren, 1997), which can enable teachers to conduct the evaluation activities in a short time. Peer evaluation engagement also

benefits in improving the students' soft skills such as critical thinking (Rohmah et al., 2021) and self-regulation (Meusen-Beekman et al., 2016). Researchers aiming to improve these peer evaluation skills via group awareness indicators from their learning logs (Kasch et al., 2021) or interactive peer evaluation platforms with backward feedback (Lin et al., 2021) emerged in recent years.

To conduct effective evaluation activities, participants need a clear impression of how they should evaluate others, in case they will just give casual ratings or compliments thus making the evaluation invalid. Instructors should give them rubrics (Andrade, 2005) and articulate evaluation criteria in a clear manner depending on different learning contexts, which has become a consensus in related research (Gueldenzoph & May, 2002). Regarding the social-emotional issue, it is known that the peer evaluator is not willing to make unfavorable judgments about the person if (s)he is exposed to peers (Cheng & Warren, 1997) because peer grading is sensitive data. Hence, to alleviate the impact of such pressure, a peer evaluation system should guarantee the privacy of evaluators and enable flexible visibility of evaluation scores depending on different contexts. From the learning analytics perspective, the re-use of these peer evaluation data is seldom discussed in the former peer evaluation platforms, hence we aim to go further on the role of peer evaluation in the data-driven ecosystem.

In parallel, peer evaluation reliability was focused on since the quality of peer evaluation remains promising (Aminu et al., 2021). Current studies on the online environment present several scaffolds to improve the reliability of peer assessment with enhanced privacy (Tharim et al., 2016) and group awareness support (Kasch et al., 2021). Nevertheless, there still remains to unbalance of grader reliability due to individual difference among learners, which lead to less accurate evaluation results in practice. To address this issue, researchers made attempts to alter the final rating values according to grader-specific variables such as previous rating tendency (Masaki, Maomi, et al., 2008) and previous grades of relevant tasks (Bjelobaba et al., 2022; Piech et al., 2013). However, the possibilities of learning model data are seldom explored to make a comprehensive prospect of these variables with existing peer evaluation designs holding limited data aggregation and re-use features (Liang, Toyokawa, et al., 2021).

Chapter 3

Group learning based on evidence (GLOBE) framework and systems

With increasing learning log data accumulated in the digital environment, research on Computer-supported Collaborative Learning (CSCL) (Stahl et al., 2006) yielded opportunities to scaffold collaborative learning with front-end information technologies and a data-rich environment. Group Learning Orchestration Based on Evidence (GLOBE) (Liang, Majumdar, & Ogata, 2021) presents a framework for AI-based collaborative learning support with data-driven approaches in a learning analytics-enhanced environment (Majumdar et al., 2019). There are four phases of collaborative learning: group formation, orchestration, evaluation, and reflection, where data flow and AI scaffold are empowered by the group formation and peer evaluation modules (see Figure 3.1). Under the GLOBE infrastructure, learning analytics for group learning like algorithmic group formation could get increasingly automated as data on group work experience grows. The following sections will introduce the two systems and the continuous data flows among the GLOBE modules.

3.1 Group formation: algorithmic grouping using logs in student model

As for the group formation module, we presented a group formation system that enables student models from different data sources underpinned by genetic algorithms and LEAF infrastructure that aggregates multiple learning logs (Ogata et al., 2018; Ogata et al., 2023).

To represent a group formation, one combination of students constructs a candidate

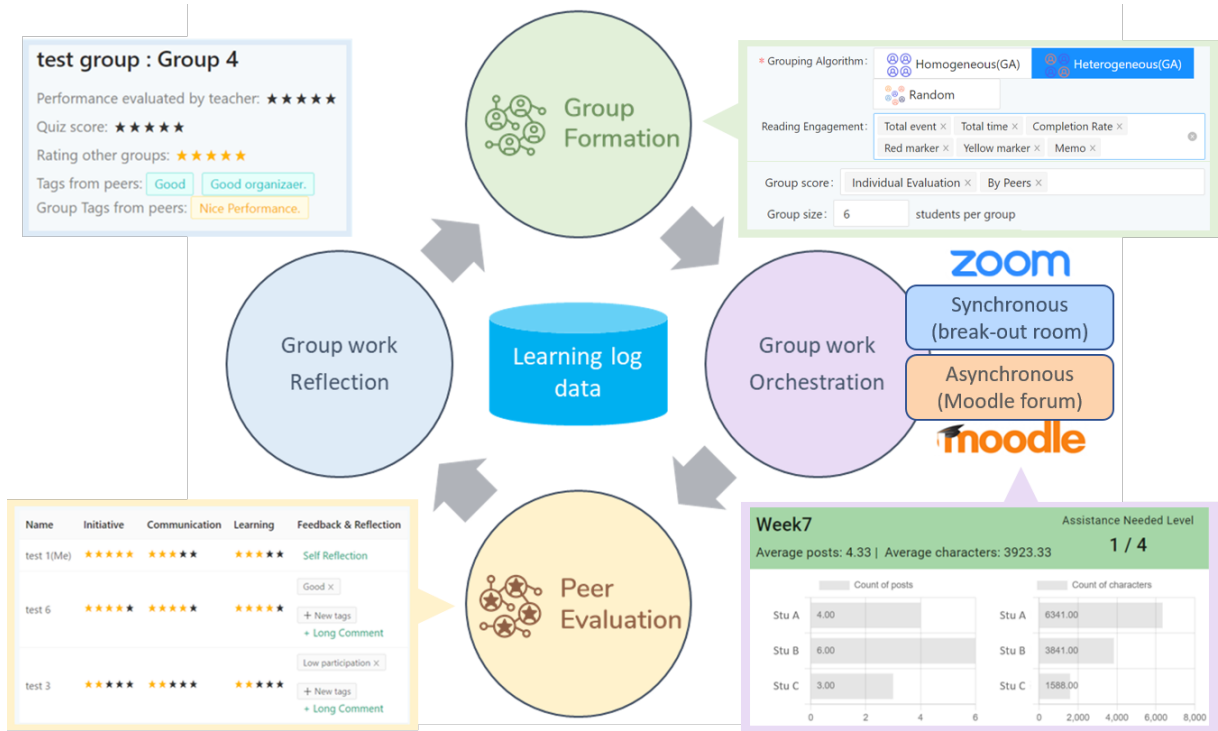


Figure 3.1: GLOBE framework and its implementation systems

individual (G) as a set of randomly-ordered students (s) partitioned by groups (Figure 3.2). 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 3.3 illustrated an example of metrics representation where each student (s) is represented by a column vector with a characteristic (c) being represented as a dimension. Figure 3.4 shows the parameter setting page and the current available input parameters from multiple data sources.

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 3.1 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 $\bar{x}_{j,g}$ 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.

$$F_g = \sum_{s=1}^S \sum_{j=1}^C (c_{j,s} - \bar{x}_{j,g})^2, F = \sum_{g=1}^N F_g \quad (3.1)$$

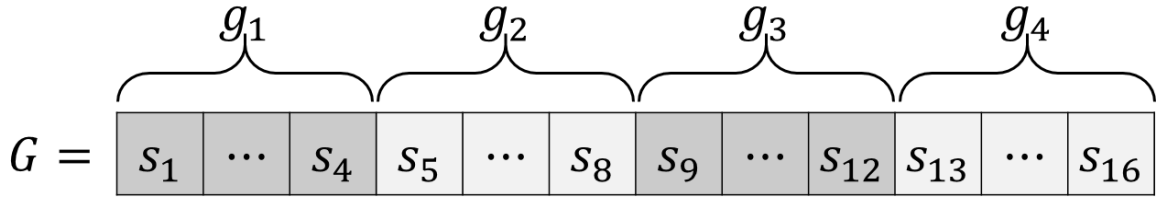


Figure 3.2: Representation of a candidate group formation as a vector of students divided into groups, illustrated by an example of 4 groups of 4 students (Flanagan et al., 2021)

Using relationship data, the algorithm enables students with good relationships (type 1) to be assigned to the same group. Conversely, the negative relationship (type 2) will be considered to separate students. Figure 3.5 shows an example of relationship data. In line with this data, student A and C, student B and G, student E and F will be given priority to be together while student C and D, E and H will be separated. Once the relationship data indicating positive and negative relations between students is uploaded, a graph shown in Figure 3.5 will be visualized. The red lines indicate pairs with poor relationships and blue lines indicate that with good relations. Each red dot represents a student and the name will be displayed with the mouse moves on it.

The relationship coefficients R_g represents how many positive or negative relationships are considered within one group by adding sub-coefficient p_g and n_g , which are dependant to the configured weight of positive (w_p) and negative (w_n) relationship (see Equation 3.2). We assume that for homogeneous algorithm $T_p = 1$ and $T_n = -1$, and for heterogeneous context $T_p = -1$ and $T_n = 1$. After taking relationship coefficient into consideration, the

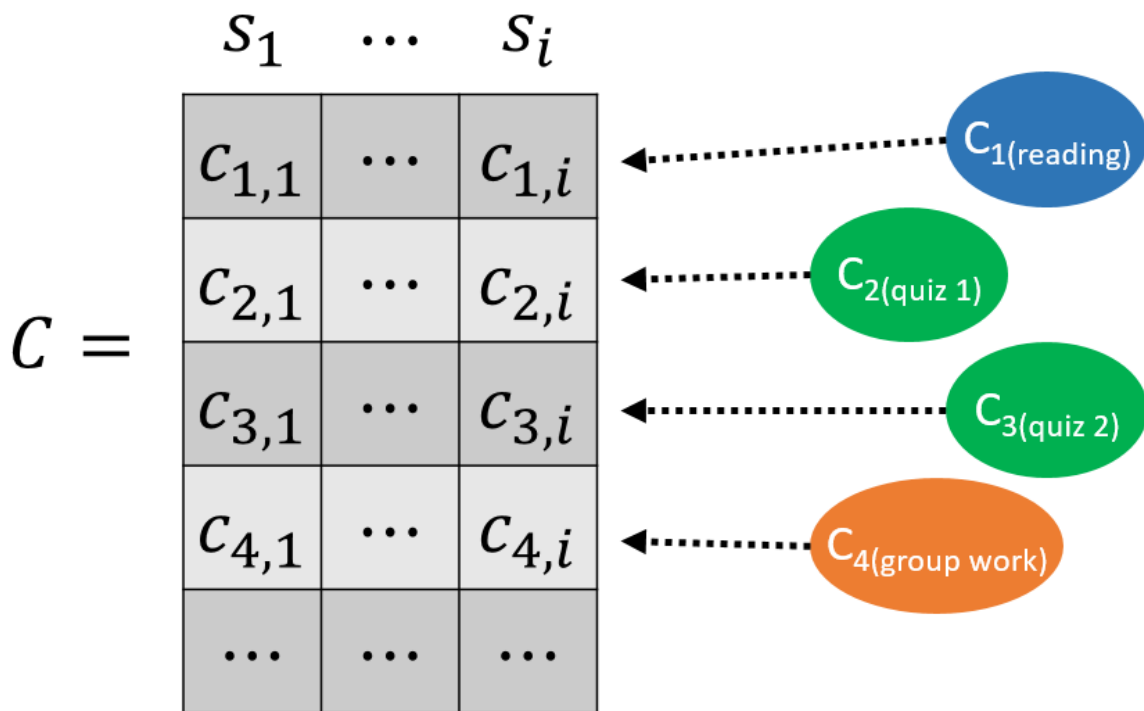


Figure 3.3: The mapping of student model variable values to the student characteristic representation matrix (Flanagan et al., 2021)

Algorithm

- Homogeneous
- Heterogeneous
- Random

Reading logs

- Times of operation
- Reading time
- Completion rate
- # of highlight markers
- # of difficulty markers
- # of memos

Quiz scores

- File upload
- Moodle quizzes
- BookRoll quizzes

Previous group work

- # of forum posts
- Interval of posts
- Sentimental analysis
- Ratings from teachers
- Peer ratings (individual)
- Peer ratings (group)

Figure 3.4: Algorithms and parameters used in the group formation module

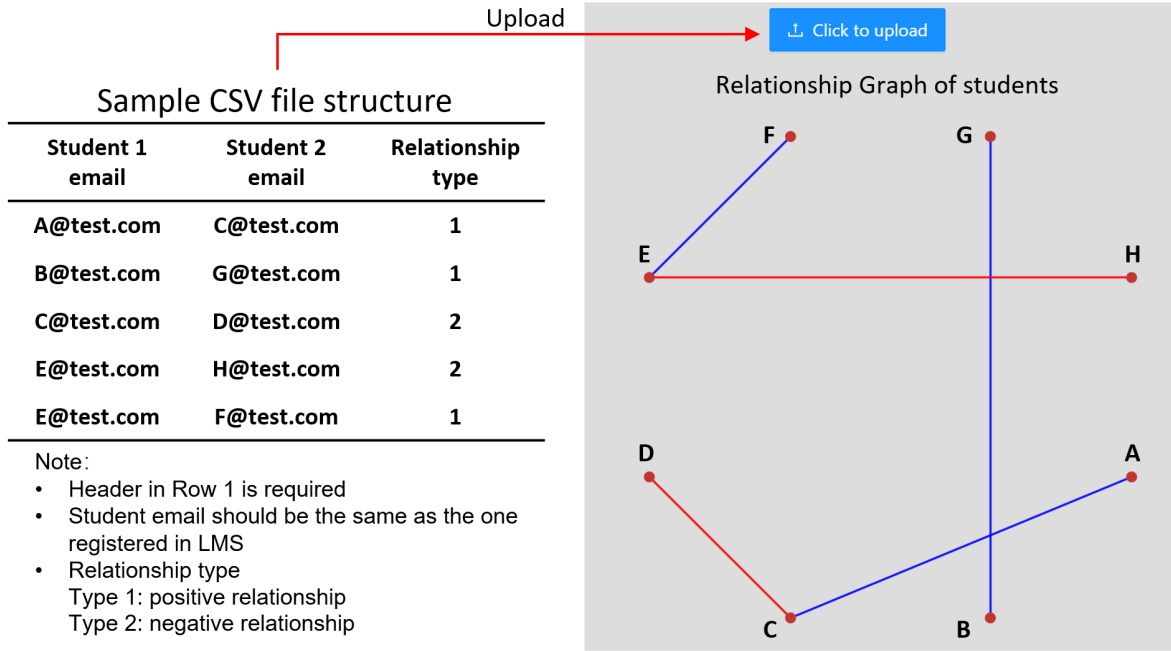


Figure 3.5: Creating and visualizing friendship for group formation.

fitness value of one evolution individual is reformed according to the weight of S_g and R_g in Equation 3.2. There remains a challenge to fine-tune the weight (λ) of new coefficients in consecutive practice.

$$R_g = T_p \sum_{p_g=1}^{P_g} w_p + T_n \sum_{n_g=1}^{N_g} w_n, F = \sum_{g=1}^G (S_g + \lambda R_g) \quad (3.2)$$

After creating groups, teachers can also check the group's homogeneity and the details of each attribute of the group members. Figure 3.6 serves as examples of a heterogeneous group and a homogeneous group formed by the system showing its F_g in the equation 4.1 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 work. For the heterogeneous grouping algorithm, we can ensure that these students 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 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.

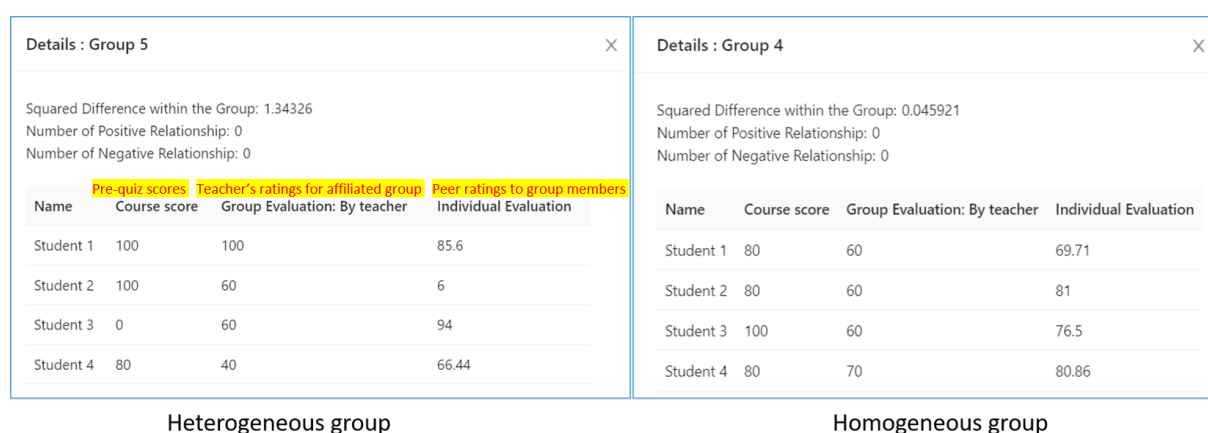


Figure 3.6: Example of the group formation details of a heterogeneous group and a homogeneous group (raw scores scaled to 0-100)

The system also provides a regrouping function to create groups for jigsaw activity, where students first form expert groups to discuss their expert section, then they go back to jigsaw groups with different group mates to present the expert section to their peers and work cooperatively to assemble a complete subject based on their findings and discussions.

The system also offers a regrouping function for facilitating the jigsaw activity (Aronson & Bridgeman, 1979). In this activity, students initially form expert groups to discuss their respective sections of expertise. Subsequently, they transition to jigsaw groups, where they collaborate with different peers to present their expert sections and collectively construct a comprehensive subject based on their research and discussions. To prevent students from being in the same group as before, the group formation system can regenerate jigsaw groups from the expert groups. This ensures a diverse mix of group mates for the collaborative task.

3.2 Group work evaluation: peer evaluation and feedback

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 (Liang, Toyokawa, et al., 2021), group members can rate other individuals in their group or another group by just clicking the stars in the interface (see Figure 3.7). 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 (see Figure 3.11). Once the ratings and comments are provided, the system shows them to the specific users with real-time ratings and textual feedback without association with the evaluator's name (see Figure 3.8). 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.

3.3 Supporting continuous data-driven group works under GLOBE

The continuous data-driven support provided throughout the two phases of GLOBE is summarized in Figure 3.10. The data collected by the GLOBE system is utilized cyclically by both phases. A simple randomized grouping followed by the use of evaluation scores for subsequent grouping provides a feasible solution to the cold start problem in data-driven research (van der Velde et al., 2021). As shown in the figure, the peer and teacher evaluations are logged into the learning record store as part of the student model (orange circles) and 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. At-risk students and groups are shown in red in Figure 3.9, indicating that they need more attention from the instructor.

In addition, during the evaluation phase, the student model attributes used in the

Rubrics				
	(1)悪い理由	(2)改善案	プレゼン	
1	3つの画像がある	改善案を述べている	プレゼン	プレゼンを行なった
2	デザイン原理、ユーザー体験目標、ユーザビリティ目標のいずれか1つに言及して考察している	いずれか1つ増えるごとに+1点 <ul style="list-style-type: none"> 改善案が具体的である 改善案が妥当であると感じられる (1)の議論を踏まえている 		
3	上記のいずれか1つに言及して考察している			
4	上記の全てに言及して考察している			
Input ratings				
名前	(1)悪い理由	(2)改善案	プレゼン	コメント/感想
+ [Redacted]	★★★★★	★★★★★	★★★★★	+新しい +長いコメント
+ [Redacted]	★★★★★	★★★★★	★★★★★	自己評価
+ [Redacted]	★★★★★	★★★★★	★★★★★	+新しい +長いコメント

Figure 3.7: Interface of peer rating with three criteria set by the teacher

group formation phase can be utilized as performance indicators to determine the reliability of each evaluator's peer ratings, as suggested by (Piech et al., 2013), addressing the impact of biased peer scores. Raters with higher scores in the group formation indicators are modeled as high-reliability students and given a higher weight when calculating the scores for an individual using weighted average scores. The estimated weight is referred to as rater potential (P) (Liang, Gorham, et al., 2022). This function has been visualized in the system, as shown in Figure 3.11, displaying both the raw score and the weighted score considering the reliability of each rater. The consistency indicators (Fukazawa, 2010) describing the agreement with instructor-assigned grades (validity, V) and average student-assigned grades (reliability, R) are presented for the instructor and stored as learner model data for further learning analytics purposes in subsequent learning activities.

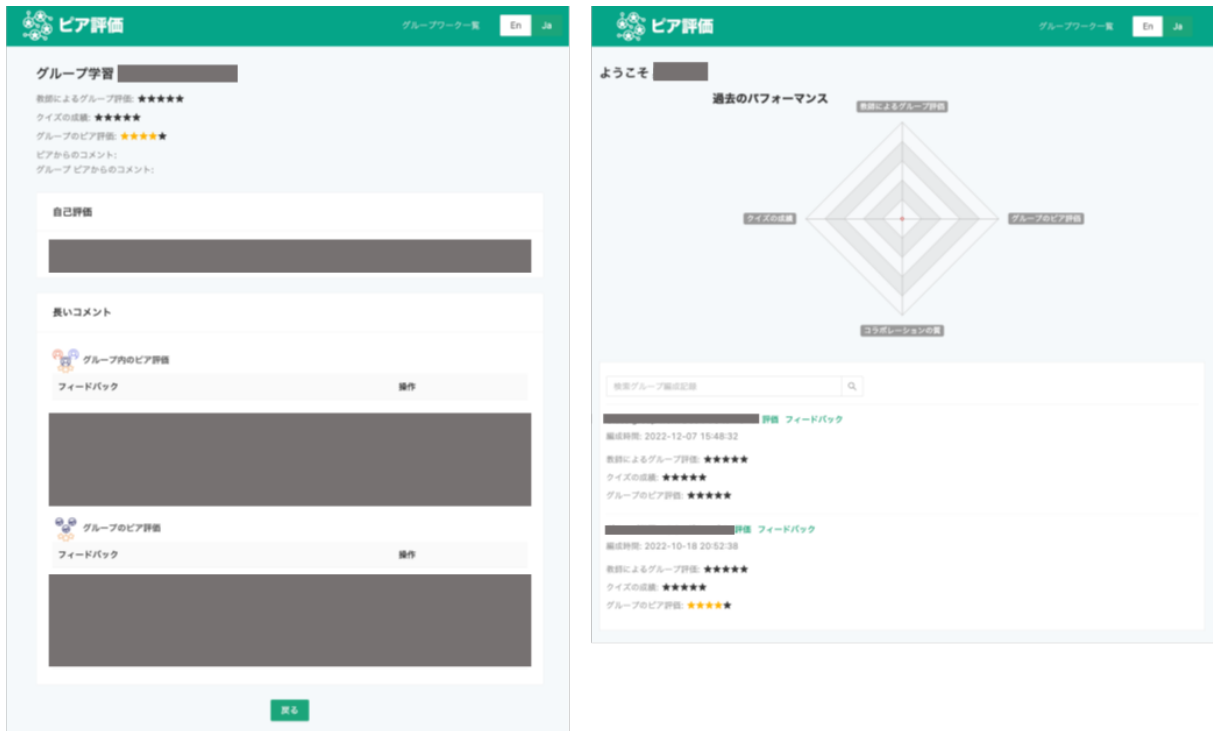


Figure 3.8: Interface of peer feedback panel



Figure 3.9: Interface of peer rating with three criteria set by the teacher

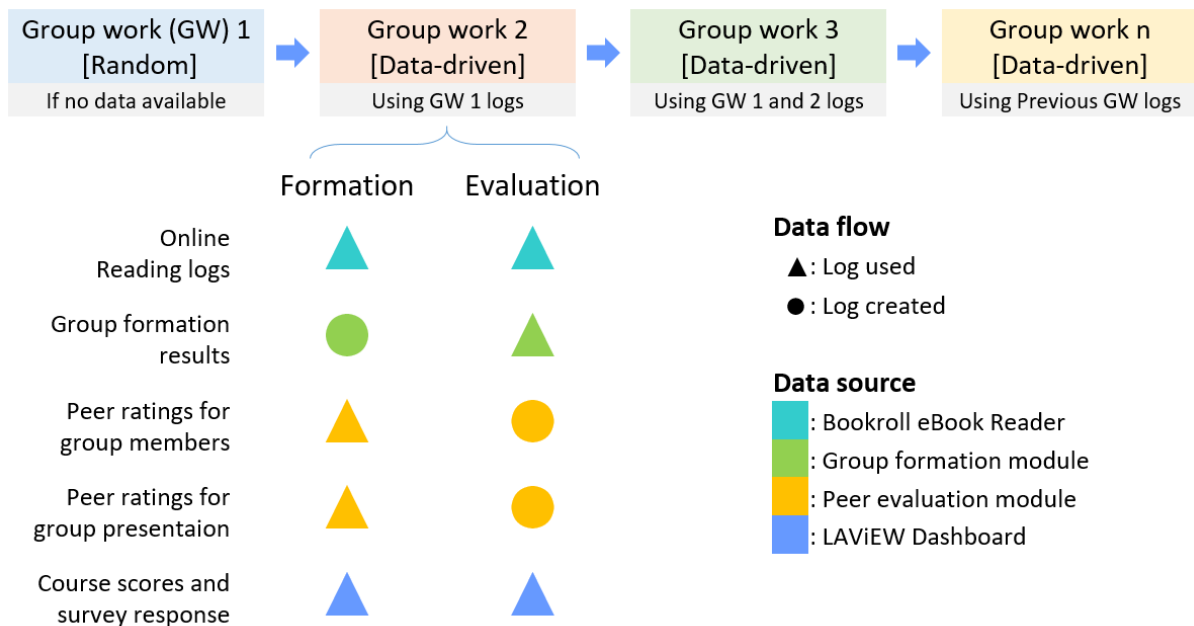


Figure 3.10: Example of a continuous data-driven support data flow under GLOBE

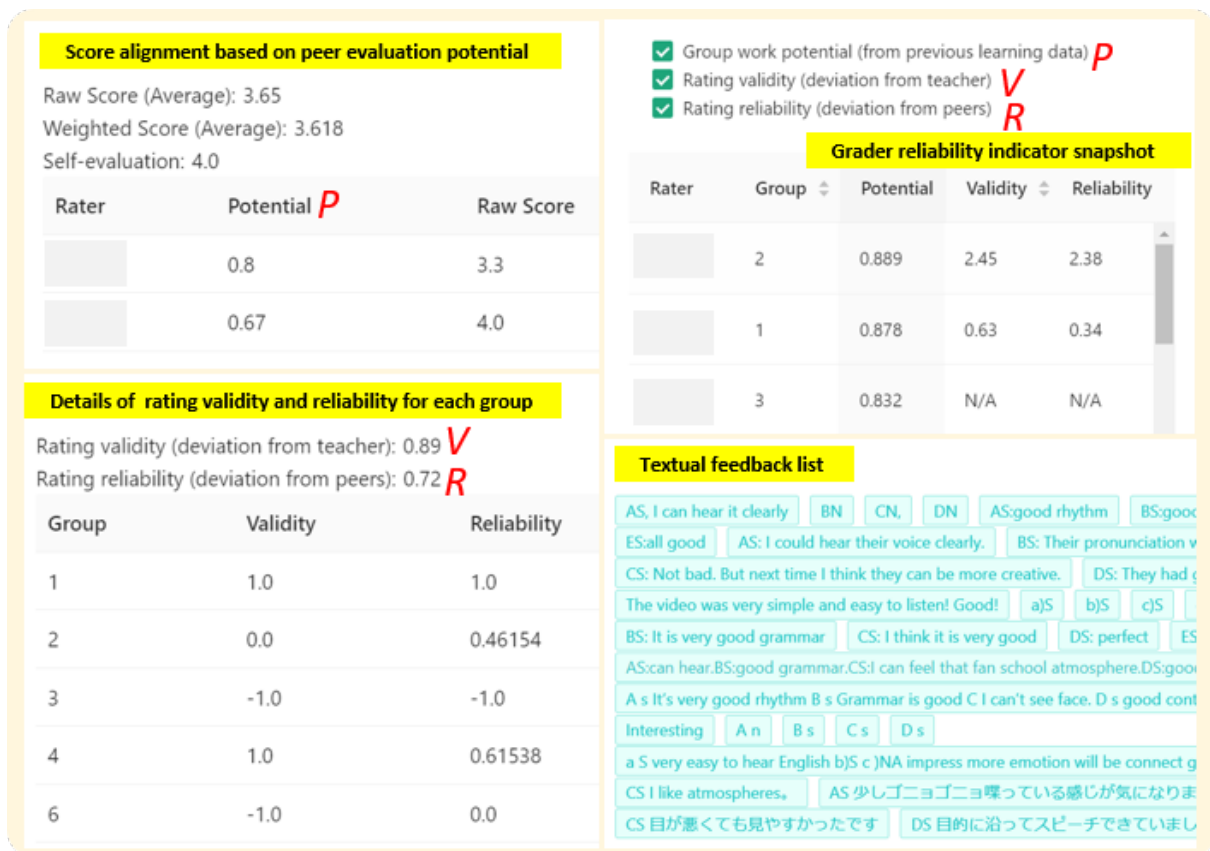


Figure 3.11: Example of a visualization of the weighted score considering the reliability of peer ratings

Chapter 4

Synthesize: Group formation using learning log data

To establish a comprehensive data-driven group learning support ecosystem, the initial stage involves synthesizing learning log data from various platforms to facilitate group formation. This chapter presents two studies on group formation. The first study demonstrates the utilization of score data to create groups for jigsaw activities in primary school math classes, illustrating an example of group formation through numerical data. Additionally, the second study proposes a solution to incorporate annotation attributes within the context of active reading for group formation. This is achieved by dividing reading markers into clusters based on the similarity of the marker text and is supported by an empirical study conducted in a middle school English reading class.

In both Study 1 and Study 2, a heterogeneous grouping strategy is employed. These studies specifically investigate learning in school settings through peer-assisted activities, recognizing the significance of knowledge construction among peers (Fischer et al., 2002). Following the principles of the ZPD (Zone of Proximal Development) theory (Vygotsky, 1980), heterogeneous groups are recommended for such contexts. In this context, heterogeneity within a group refers to the varying levels of prior knowledge and cognitive skills exhibited by group members. These attributes are estimated based on factors such as course grades, communication skills, and unfamiliar words when reading an article.

4.1 Study 1: Group formation based on knowledge and relationship

4.1.1 Aim and research question

Considering the challenges faced by teachers when implementing group work, as depicted in Figure 1.1, the need for reliable support in executing and managing group formation activities promptly and effectively becomes paramount. In the previous chapter, we introduced a system that addresses this need by providing teachers with assistance in group formation and analytics, leveraging learner model data (Boticki et al., 2019). In this study, we implemented the system to aid teachers in conducting group-based classroom activities within an actual school setting. Our investigation focuses on evaluating the system’s effectiveness by examining its primary impact on student engagement and affective states. To compare the outcomes of group work facilitated by teacher-formed groups and computer-formed groups, we conducted practical experiments. The specific research questions guiding our study are as follows:

RQ1. How do the computer-formed groups affect the students’ engagement in in-class group work?

RQ2. How do the computer-formed groups affect the students’ affective states during in-class group work?

4.1.2 Learning Context and participants

The study was conducted in a primary school maths problem-solving class covering several topics. For two different classes, two different teachers conducted the class respectively but the topic is the same. Two classes first underwent activities 1 to 3 with teacher-formed groups as baseline conditions. Then, the group formation was changed and done according to the system, and activities 4 to 7 were conducted as an experiment class. Each class is of the same length and the topics are in the same order. It maintains that data from each class are comparable. The main data for analysis of the research comes from the voice records throughout the class via USB headsets and microphones. In total, 13462 pieces of voice data that cover text and affective scores (6030 pieces for class 1 and 12767 pieces for class 2) were collected. After data cleaning, the data for analysis covers 7 lecture topics of 11 in-class activities (See Table 4.1, TG means groups formed by the teacher, and CG means groups formed by computer).

The experiment was conducted in a primary school in two Grade 5 classes. There are 32 students in 12 groups for class 1 and 33 students in 12 groups for class 2. However, not all of the 65 students participated all the class due to uncontrollable issues.

Table 4.1: Summary of data collection.

Activity	Topic	Type	# of students	# of voice data
A1 (Initialization)	Square	TG	32	1611
A2 (Trial)	Square	TG	65	1480
A3 (Baseline)	Multiplication	TG	65	2325
A4 (Trial)	Quantity per unit	CG	65	5005
A5 (Intervention)	Percentage	CG	65	3041
A6	Percentage	CG	33	2084
A7	Percentage	CG	33	2223

4.1.3 Research Design

To make a comparison between groups formed by the teacher and by the system, we adopted a within-subjects design (A-B design). We conduct the study with a single cohort of primary school students in Grade 5, however, the indicators observed are at a group level that keeps changing based on teacher-generated and computer-generated grouping, the A and B conditions. Activity A2 and A4 is the first attempt for each condition, to reduce the novice effect, we choose activity A3 (applied problems of multiplication) and A5 (applied problems of percentage) for both Class 1 and 2 for the data analysis in this research. We assume activities A3 and A5 are similar and comparable since both of them focus on math problem-solving in similar topics.

4.1.4 Procedure

The in-class group work adopts the “jigsaw learning method” (Aronson & Bridgeman, 1979) consisting of two different phases (see Figure 4.1). Each student will work in a “knowledge exploration phase” and a “knowledge exchange phase” during one class, which corresponds to two different group combinations. In the knowledge exploration phase, students work on a solution with the same idea. They discuss and check their solutions

with members within the knowledge exploration group and illustrate ideas to each other. After that, students from different knowledge exploration groups go to knowledge exchange groups and explain the idea to those who solved the problem differently. In the knowledge exchange phase, which is a knowledge exchange phase students exchange ideas and talk about different solutions.

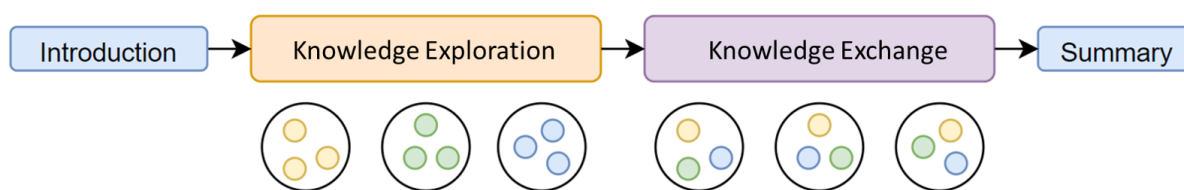


Figure 4.1: Process of the in-class group work.

Take the topic “the square of a trapezoid” as an example, the system first collects data from different sources and then forms groups accordingly. A pre-test about the estimation of triangle squares is conducted at the BookRoll system to confirm the level of understanding of these learned items. The test results are used as input parameters of the group formation. Meanwhile, course scores from the LA view dashboard indicating communication skills and performance data of previous performance scores relating to the topic “Square” are extracted to conduct group formation in the system. Besides, relationship data are created by teachers and uploaded in the tab of “relationship” on the group formation parameter setting page. In this context, the system first uses the friendship algorithm to group students with positive relationships, then groups the rest of the students using the jigsaw algorithm as illustrated in Chapter 3.

Before the class starts, the tablets and headset microphones are prepared and set in the classroom. At the commencement of the class, the teacher writes the goal of the class “Square of a trapezoid” on the blackboard and puts forward a specific problem of calculating the square of a trapezoid. The problem is to be solved throughout the class thus motivating students to learn. Then the group work activity starts and the utterances are recorded for each student respectively. For the topic of “the square of a trapezoid”, each knowledge exploration group will be asked to discuss either of the following solutions: making a parallelogram, dividing into two triangles, or dividing into a triangle and a parallelogram. And in the knowledge exchange phase, students in one group will share all three solutions with other members so that all students know the three solutions. Finally,

the teacher gives the summary of the whole class and students write down three ways of calculating the square of the trapezoid on the blackboard. After the class, a feedback seminar is conducted where teachers reflect on their teaching experience and share their doubts and feelings.

4.1.5 System Usage

In the implementation, input parameters from three data sources were considered based on related works and teachers' opinions. The heterogeneous algorithm was applied using the following parameters, and then the regrouping function is used to form jigsaw groups.

- Bookroll quiz scores: The pre-test indicating the pre-knowledge of the learning subject was done on the online textbook Bookroll using its quiz function and the quiz scores are acquired as an important input source of the group formation.
- Course skill scores: Communication skills, way of thinking, and academic skills are provided as scores by teachers and uploaded in the LA view dashboard.
- Friendship data: The friendship data indicating both positive and negative relationships of students is uploaded in the group formation tool since the teacher stressed that students with negative relationships should not be grouped together.

4.1.6 Data Collection

For the utterance data indicating students' engagement, the duration of each speaking was recorded and then the speech data was textualized by speech-to-text API. We divided the text into tokens (meaningful words) by Node.js TinySegmenter API for Japanese tokenization (Kudo, 2016). Then the words are counted as the number of tokens. The teachers' speech data was filtered before the analysis as well.

The affective scores data indicating affective states are transformed from utterance data as well by pattern recognition API. Four affective states: joy, vitality, anger, and calmness, were computed into scores for each piece of utterance. Joy indicates the student works in a positive mood. Vitality denotes how actively the student performs in group work. Anger implies conflict within group members. Calmness represents low engagement and low motivation. Each affective score is standardized into the range of 0 to 1 before analysis.

4.1.7 Data Analysis

To explore the difference in the knowledge exchange phase between teacher-formed groups and computer-formed groups and answer research question 1, we do analysis at both the group level and individual level. Comparing the overall mean provided a group-level aggregation of engagement, we look into the effect of intervention condition (CG) in three indicators: times of utterance, duration of utterance, and the number of tokens. Since the data of the three indicators do not satisfy the normal distribution according to the Shapiro–Wilk test ($p < 0.05$) (Shapiro & Wilk, 1965). We adopt non-parametric tests to measure the significance of the difference. Mann-Whitney U test is conducted and the effect size is calculated respectively for the three engagement indicators.

Further analysis was done to understand transitions of cohorts of specific engaged students within phases of one activity or across activities. Individual learners' engagement category, based on their speaking duration, was considered to do this analysis. The transitions in engagement categories were looked at from two different perspectives. One perspective is between two activities for each phase and overall. Such analysis was afforded by the iSAT tool which could visualize transition patterns across phases with SAT Diagram (Majumdar & Iyer, 2016).

The affective scores of two independent samples are compared by independent t-test to answer research question 2. Since the Shapiro–Wilk test of affective indicators ($p = 0.053 > 0.05$ for anger, $p = 0.299 > 0.05$ for calmness, and $p = 0.511 > 0.05$ for joy) shows normal distribution except vitality ($p < 0.05$), an independent T-test is done on three indicators and a Mann-Whitney U test to vitality score. The null hypothesis establishes that the means of the affective scores are of equivalence, and correspondingly the alternative hypothesis establishes that the means are of difference.

4.1.8 Result and Inferences

Engagement

Knowledge Exchange Phase As is shown in Table 4.2, all three indicators of group work engagement suggest significant improvement in the intervention condition (CG). Groups formed by the system have more times of utterance ($M = 110.4$, $SD = 58.85$), longer utterance duration ($M = 734.02$, $SD = 375.52$) and also more meaningful tokens ($M = 609$, $SD = 340.89$) in comparison with groups formed by teachers: T ($M = 35.46$, SD

= 13.98); D (M = 230.11, SD = 108.03); N (M = 256.38, SD = 89.33). The effective size of the three indicators are 0.452, 0.405 and 0.437 respectively, which indicates a medium to large effect (Cohen, 1988).

Table 4.2: Difference in engagement indicators for knowledge exchange phase.

		N	Mean	SD	p	Effect size
Times of utterance (T)	TG	24	35.46	13.98	0.000***	0.452
	CG	20	110.4	58.85		
Duration of utterance (D)	TG	24	230.11	108.03	0.000***	0.405
	CG	20	734.02	375.52		
Number of tokens (N)	TG	24	256.38	89.33	0.000***	0.437
	CG	20	609	340.89		

***p < .001.

Figure 4.2 shows the transition graph of the utterance duration indicator in the knowledge exchange phase between two conditions. In the transition graph, three strata (Top, Mid, and Low) are defined for each phase independently and presented in Table 4.3. The Top-Mid cutoff is delimited using mean plus standard deviation and Mid-low cutoff by mean minus standard deviation. NP (Not-participate) layer indicates absence in this phase. We can see more students start to participate in discussions in computer-formed groups since the transition from NP to Top and Mid account for 19% for the knowledge exchange phase. Meanwhile, computer-formed groups encourage active students to even speak more than the baseline condition. It is indicated that more students' utterance duration reaches a high level in A5 activity which is based on computer-formed groups.

Table 4.3: Cutoff of three strata of the utterance duration transition graph.

Strata	A3-knowledge exchange	A5-knowledge exchange
Top-mid cutoff	194.25 seconds	492.80 seconds
Mid-low cutoff	62.62 seconds	50.92 seconds

Knowledge Exploration Phase Table 4.4 shows the result of the Mann-Whitney U test for idea exploration group work on this regrouping activity at the group level. Con-

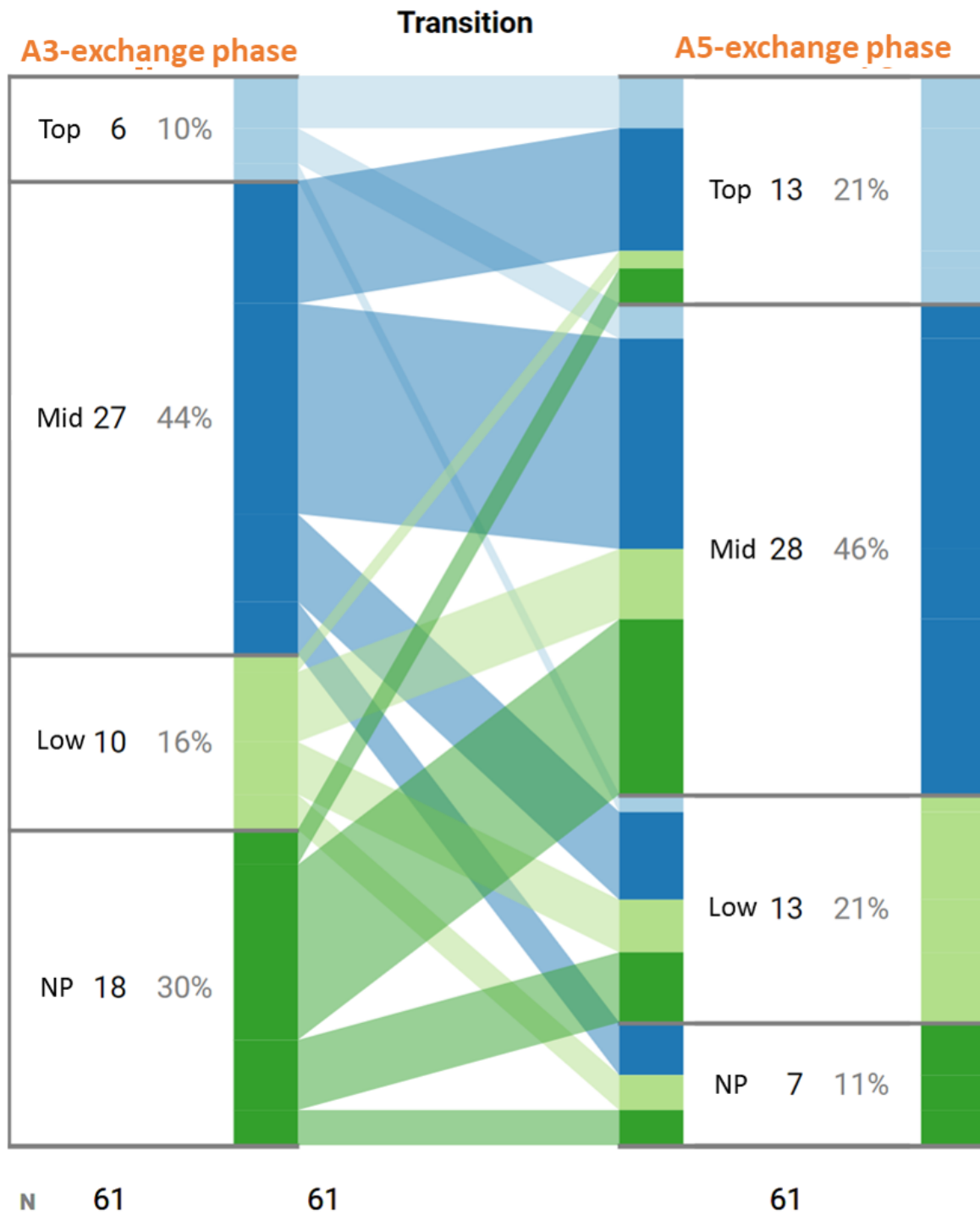


Figure 4.2: Transition patterns of utterance duration in knowledge exchange phase between activity A3 and A5.

verse to the knowledge exchange phase, it is indicated that for the engagement indicators, teacher-formed groups perform better in this context in all three indicators with small

effect sizes of 0.319, 0.322, and 0.303 respectively.

Table 4.4: Difference in engagement indicators for knowledge exploration phase activity.

		N	Mean	SD	p	Effect size
Times of utterance (T)	TG	18	53.94	24.08	0.000***	0.319
	CG	20	26.6	24.89		
Duration of utterance (D)	TG	18	352.77	135.41	0.000***	0.322
	CG	20	177.69	132.49		
Number of tokens (N)	TG	18	312.17	203.80	0.000***	0.303
	CG	20	136.5	135.89		

***p < .001.

A simple observation of transition of the duration of utterance is also implemented in the reshuffled group as is shown in 4.3. We found that still, 15% of students from Mid, Low and, NP layers in teacher-formed groups come to Top layer in knowledge exploration activity, which makes the percentage for Top layer increase in computer-formed groups. However, 13% of students in Mid layer kept silent without any utterance in the computer-formed groups.

Affective States

Figure 4.4 depicts the result of the test on the affective scores at the group level and the mean of each standardized effective score for each group is labeled on the bars. As is indicated in the figure, the joy and vitality affection present the same pattern that the experiment class where groups are formed by the system has a higher score of these positive affections. On the contrary, regarding negative affections, calmness, and anger denote the opposite result, with the control group having higher scores. However, only the difference in joy proves to be at a significant level in the statistics ($t(24)=0.004 > 0.05$) and the null hypothesis is rejected. For calmness ($t(24)=0.143 < 0.05$), anger ($t(24) = 0.777 > 0.05$), and vitality ($p=0.066 < 0.05$, effect size=0.079, indicating very low effect), the null hypothesis cannot be rejected within a confidence level.

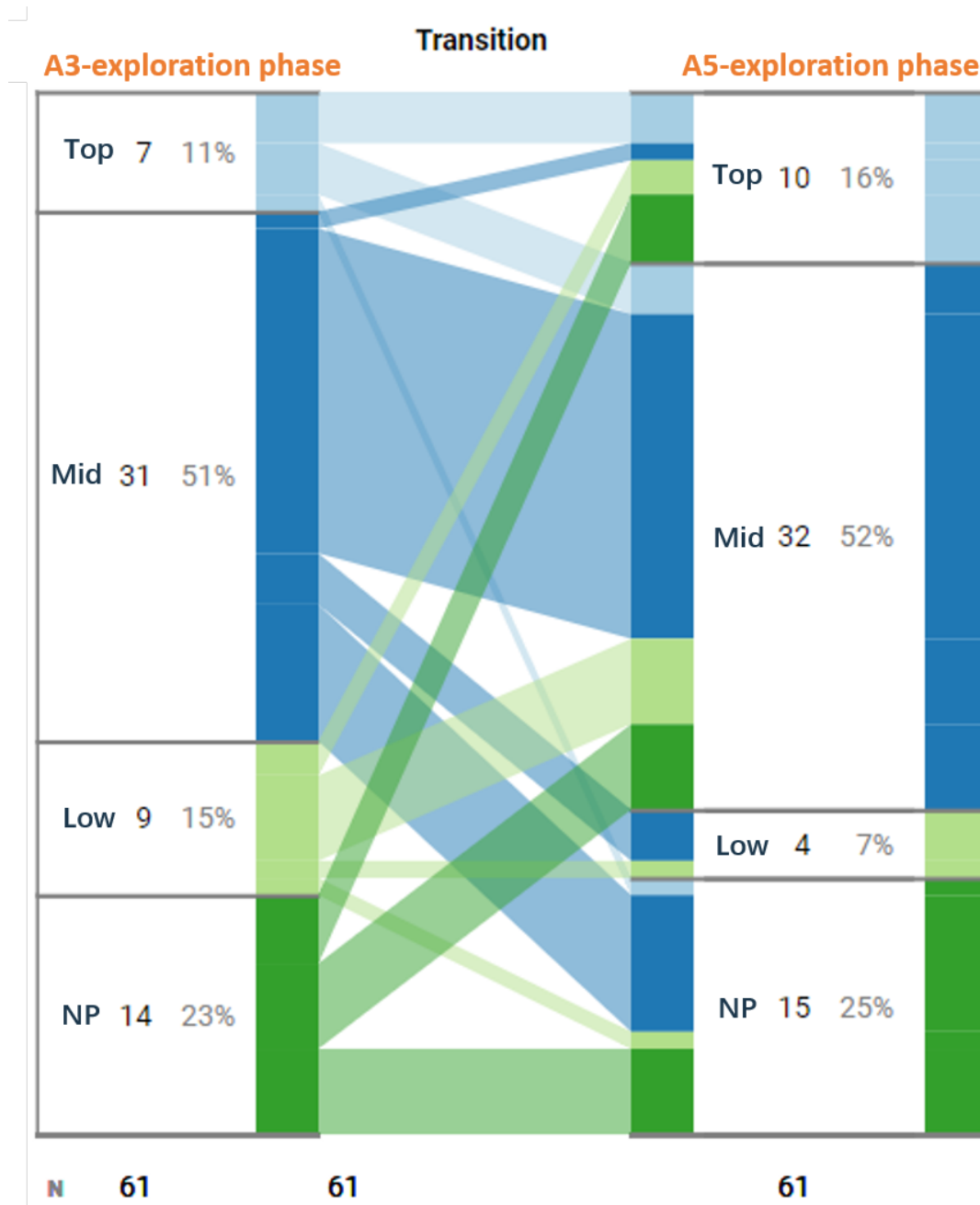


Figure 4.3: Transition patterns of utterance duration for knowledge exploration phase activity between activity A3 and A5.

4.1.9 Discussion

RQ1: How do the computer-formed groups affect the students' engagement in in-class group work?

The results show the difference in the process of the group work between groups formed by teachers' experience and by evidence data using the system. Generally, each group speaks

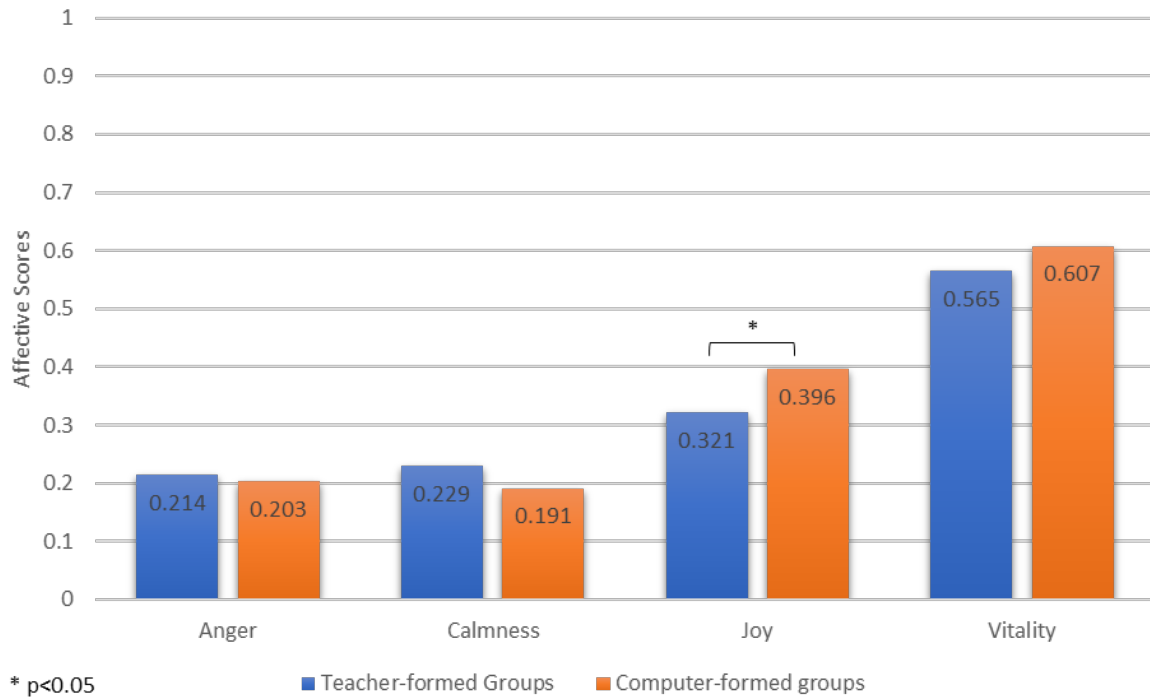


Figure 4.4: Difference in affective state scores for knowledge exchange phase.

more, and the duration of utterances increases in the computer-based groups. This finding supports the superiority of the system for idea exchange activity to arouse motivation and facilitate the engagement of students. The parameters for group formation may be a key factor that determines this phenomenon. That is to say, the diversity of communication skills, pre-knowledge of the learning topic, and previous academic performance catalyze the atmosphere and facilitate interaction for idea exchange within heterogeneous groups. It is also grounded in the research in the area of the Zone of Proximal Development (ZPD) and potentially promotes the construction of knowledge and an elevated level of the mutual understanding of a topic (Nyikos & Hashimoto, 1997). The finding also agrees with the recent work that presents the effectiveness of heterogeneity of the student cohort in workshop group activities (Sivaloganathan et al., 2020). Besides, we can see that the difference reverses in the reshuffled groups for the knowledge exploration phase activity. On the one hand, it supports the effectiveness of the system and parameter settings in the knowledge exchange condition. On the other hand, we cannot deny the fact that the system is still short of flexibility in the regrouping context.

In terms of the transition graph, we can infer that the new combination of group members encourages active students to even speak more and in turn facilitate low-performance

students to participate. Even for the reshuffled group in a regrouping context, the percentage of top-level students increases in the computer-formed groups, which can be partially attributed to the work of friendship data.

RQ2: How do the computer-formed groups affect the students' affective states during in-class group work?

As for affective states, students act more positively in the groups formed by the system where their utterances showed more positive affective states such as joy and vitality. Also, students performed less reserved and less irritated in the experiment groups as is indicated in the scores of calmness and anger. The difference in joy affection reaches a significant level, we can infer that the computer-formed groups bring about more happiness for students, thus promoting the initiative of utterance and high engagement in the group work. According to the teachers' feedback, it is indicated that the novelty of the new group combination motivates students to speak more and participate more actively. We can also conclude that the friendship-priority grouping strategy utilizing friendship data reduces the conflict among group members because trust relationship and the group's willingness to handle group work challenges was positively related to individual students' group work self-efficacy (Du et al., 2019). However, since the difference in vitality, calmness, and anger do not reach a significant level, the effect of the new group composition on these affections is limited.

Implication for Teaching

Due to the busy schedule of the teachers, an informal interview with them was conducted to gather feedback after they used the system. The overall impression was positive. Teachers mentioned that unexpected combinations of students which broke the teachers' prototypes were discovered. Furthermore, teachers found new qualities about students and some students demonstrated leadership which is not found in ordinary classes, though they still have some doubts and as well. Nevertheless, there is a possibility that the parameters provided are not enough or not suitable for all the contexts of group formation. Therefore, it is imperative to discuss implementation potentials in further context.

The system can be applied to broader pedagogical scenarios where teachers can use the tool. For example, the system can support more complicated group work activities like multi-phase in-class regrouping activities beyond the one illustrated in this study (Figure

4.5). Before the class, the teacher can assign an online pre-test to students and then form groups based on prior knowledge indicated in the test. Since the system can form groups in seconds, it is convenient for teachers to create groups just in class for different phases of activity for multi-phase activities, even utilizing the performance data of the previous phase. The workflow can be applied not only in maths problem-solving but to other forms of collaborative problem-solving (CPS)(Pöysä-Tarhonen et al., 2018).

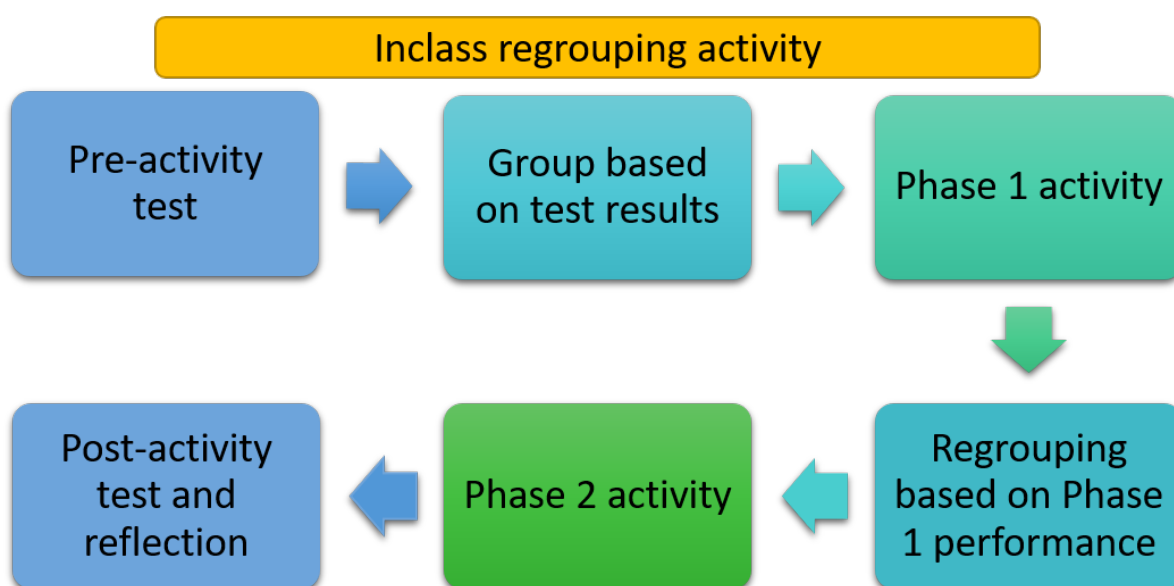


Figure 4.5: Typical workflow for activities involving regrouping.

Flipped reading is another example. Using learning logs from reading behaviors and records from LRS, the teacher can conduct flipped reading classes using the system (Figure 4.6). Since rich learning logs indicate the reading skills and preferences of students, the integration of reading data makes it easy for teachers to generate homogeneous or heterogeneous groups using data regarding reading logs. The teacher can group students with similar reading habits or preferences within the group work. During the class, there can be multiple collaborative reading activities such as kit-build concept map (Hirashima et al., 2015), peer help of reading comprehension, and topic-based collaborative writing (Bremner, 2010).

Limitation

Some limitations are identified in the present study for consideration. Regarding the system development, the reshuffle method proved to be of low performance for regrouping

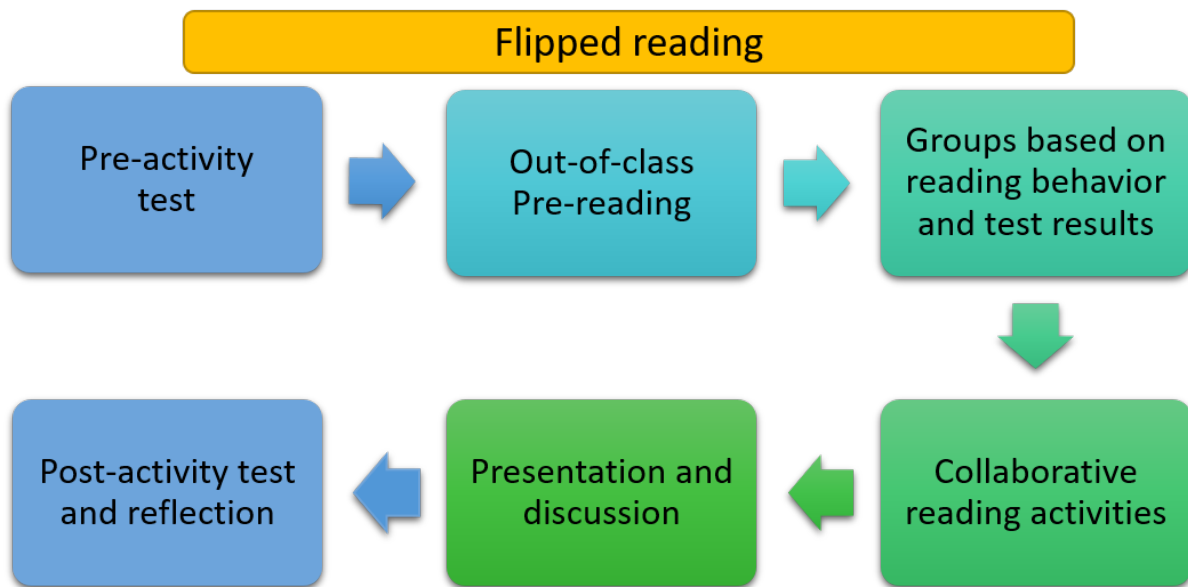


Figure 4.6: Typical workflow for flipped reading activities.

activities, which calls for improvement using different strategies. As for the experiment design, the learning topic is not perfectly identical, so the result may be affected by the topic of the class activity. Some students didn't speak even a word through the whole class or across an activity phase, which makes it hard to explain the results. Case studies may be necessary to inspect the reason behind their silence.

Besides, the precision of the transition from voice collected in class to textual data (textualized data divided by all entries of utterance record) is between 40% to 50%, which limits the deeper analysis of the specific content of the utterance. With the available data, we conducted a basic analysis of the sound features to get an initial indicator of the participant's motivation and engagement in the learning activity. However, the pattern recognition API directly coded the emotions and did not require tokenized words from the speech. Anyways in our specific context, the words were mostly limited to nouns and digits. This restricted further semantic analysis of the utterances in our current study. To make further investigation of speech signals, not only the overall duration of speech but also the spurts (Smith et al., 2016), defined as regions of uninterrupted speech, should be considered for deeper analysis. Also, more synchronized multi-modal signals are expected to catch more accurate features. For instance, the Collaboration Literacy Feedback framework including body posture and facial features provides an instructive reference for related research (Y. Kim et al., 2020). As for the interview for teachers, we

could only conduct an informal one over a group video call online due to time and access limitations to directly contact them during the period of the pandemic. Finding reasons related to ease of use by the teachers deserve further investigation, which is part of our future agenda.

For the evaluation module, which is not used in this experiment, we adopted the group assessment that only relies on the teacher’s assessment. The disadvantage is obvious it is hard to track each member’s contribution and real-time performance, thus causing social loafing and free riding. The trivial way for teachers to grade the performance group by group is not user-friendly enough. A combination of teacher evaluation and peer evaluation will provide a solution that is recommended as other researchers work (Forsell et al., 2020).

4.1.10 Conclusion and Future Work

The paper provides a feasible solution to conducting in-class group work by helping teachers divide students into groups efficiently for better group work performance. It makes an instructive technical contribution to the research on group work support systems in the CSCL field as well. An experiment to primarily test its performance was conducted as a scientific investigation, thus providing empirical evidence for the practice of CSCL systems. By using its visualization support, teachers can compare students’ performance in group work and make more informed group formation decisions in their subsequent learning designs. Compared to related work, the system proves novelty in that it integrates multiple algorithms into one same system and combines data from multiple sources that are synchronized with that system, which is designed for application in multiple contexts.

In the following studies, the implementation of the system will be extended to different activities and contexts such as regrouping activities and flipped reading mentioned in the discussion part. Besides in-class practice in school, contexts like university courses and remote education level a field for group work researches as well. Meanwhile, a more intelligent reshuffling method will be imported to enhance the flexibility to more contexts. As is pointed out in the first chapter, the research for group work support is not only confined to group formation but also covers orchestration, evaluation, and reflection. As for the orchestration phase, real-time evaluation based on speech-to-text API is under investigation. To evaluate the performance of the group work, a system that combines the teachers’ evaluation and peer evaluation is on schedule. Furthermore, in the

reflection phase, utilizing the accumulating group formation and performance data, group work analytics, and machine learning for optimized algorithm recommendation becomes possible.

4.2 Study 2: Group formation using reading annotations

4.2.1 Aim and research questions

The characteristics of participants for algorithmic group formation emanate from student model data (Brusilovsky et al., 2015) generated from the LEAF repository. Instructors can choose appropriate attributes as inputs depending on their purposes of group work (see Figure 4.7). Each selected student model attribute is quantified as numerical values, representing one dimension in the student characteristic vector shown in Figure 3.3 for fitness calculation.

When it comes to active reading, learning logs are recorded when students read using BookRoll, an e-book reader available on devices with web browsers from anywhere and anytime (Ogata et al., 2015). These logs cover reading time, completion rate, the number of markers, and memos. Related studies mentioned their connection to desirable learning outcomes (Boticki et al., 2019; Chen et al., 2021). As a data-rich environment for active reading, BookRoll provides the students' reading attributes for group formation.

However, the plain indicators mentioned above are inadequate to describe active reading behaviors, since they mainly focus on reading engagement but neglect the content level attributes. From the "record" perspective of active reading, we can extract more information from annotation patterns by investigating marking behaviors. In the BookRoll system, markers come in two different colors: yellow markers can represent unknown words, expressions, or something that is hard to understand, while red markers can denote something important. These reading annotation features are meaningful in educational practice (Toyokawa et al., 2023). For instance, common highlight markers among readers show common reading interest, while heterogeneous distributions of difficulty markers with difficulties indicate unbalanced knowledge. The existing group formation system cannot deal with the interactive features directly, which forms the gap between the current data structure and the new requirement of utilizing these marker attributes in an active reading context.

Based on the former background, this study will introduce a group formation approach using reading annotation data under an existing data-rich environment. Meanwhile, we aim to confirm the priority of this approach compared to the traditional engagement attributes-based method and random allocation. Therefore, we explored the impact of

* Group work name:

* Grouping Algorithm: Homogeneous(GA) Heterogeneous(GA) Random

BookRoll reading behaviors

Reading Engagement:

Active Reading:

Use data of all BookRoll content: Off On

Previous scores

Course score:

Moodle quiz:

BookRoll quiz:

Group work attributes

Forum:

Received ratings:

Group size: students per group (Number of students: 40 (11 Active))

This study

Figure 4.7: Student attributes from assorted resources for algorithmic group formation (Liang, Majumdar, & Ogata, 2021)

marker-based heterogeneous groups on language learning achievement and group work perceptions in an empirical study. The research questions are as follows.

RQ1: How did achievement of language learning differ among marker-based heterogeneous grouping, only reading engagement attributes-based (e.g. reading time, completion rate, etc.) grouping, and random allocation?

RQ2: How did group work perception differ among marker-based heterogeneous grouping, only reading engagement attributes-based (e.g. reading time, completion rate, etc.) grouping, and random allocation?

4.2.2 System innovation: Reading marker attributes for group formation

To feature marker attributes in the active reading context for group formation, we put forward a solution to divide reading markers into clusters based on the string similarity of the marker text. The string features fit the reading in the language learning scenarios more since the markers indicating highlight and difficulty are usually from meaningful textual contents (Ball et al., 2009). A resembling method based on marker-text similarity was applied in Chang et al. (2017). In the GLOBE system, each cluster represents one student characteristic in the student character matrix, which acts as the input of the genetic algorithm (Liang, Majumdar, & Ogata, 2021). In this way, we strengthened the existing data-driven group formation system of GLOBE to incorporate marker features for collaborative learning in active reading contexts. Figure 4.8 depicts the overall workflow of the featuring process.

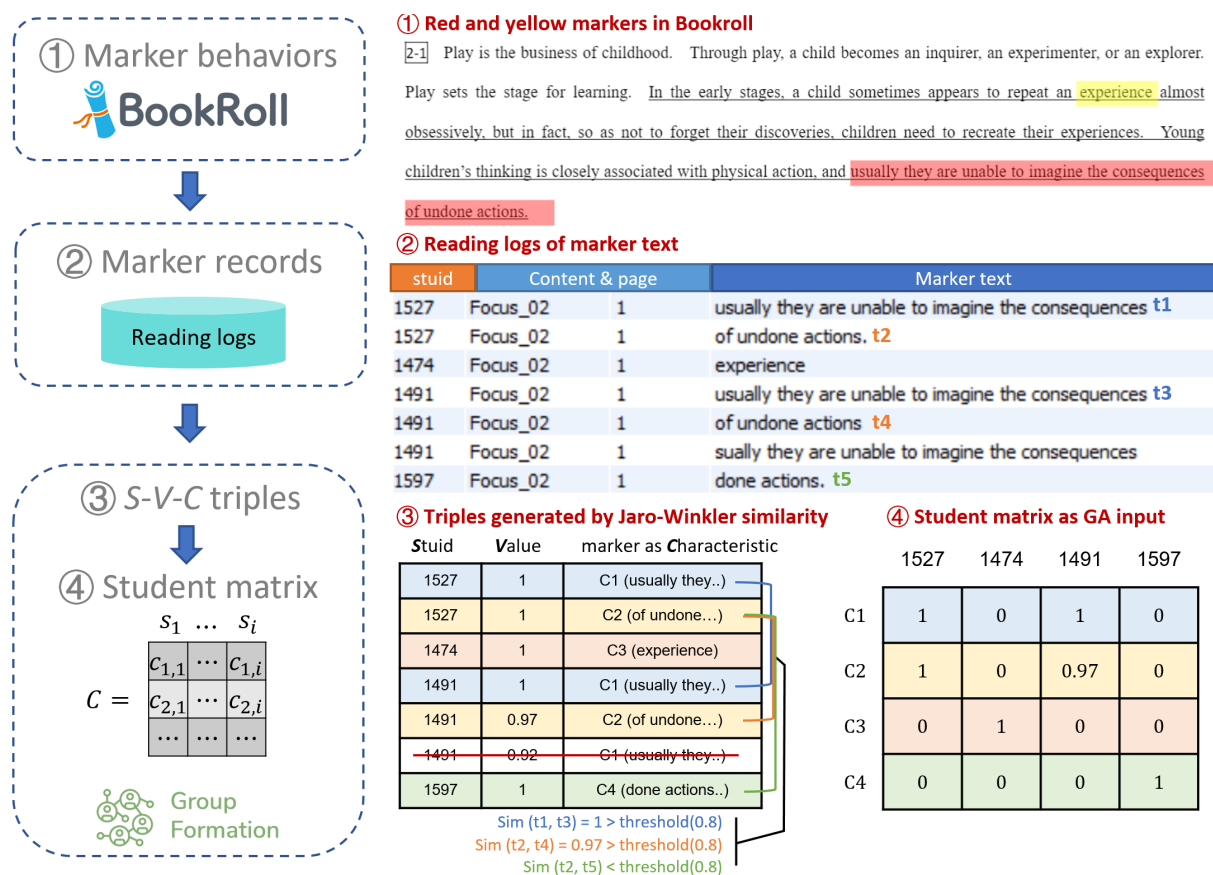


Figure 4.8: Workflow to feature marker attributes for group formation

- First, students create different markers in the E-book reading platform and each

marker is logged into the reading log database with recognized texts.

- Then, we estimate the similarity of the marker text based on the page it appears and its string features to determine common markers from different readers. In this study, we employed Jaro-Winkler similarity measurement (Winkler & Thibaudeau, 1991). Since there could be some fluctuations in the marker quality and text recognition, we do not require identical marker text. Therefore we adopted a threshold of 0.8 considering the requirement of accuracy and the number of possible marker clusters in active reading activities according to Draibach and Naumann (2013).
- Next, all marker records with marker texts are selected and assigned to clusters based on text similarity, thus generating a Student-Value-Characteristic list.
- After that, duplicated markers from the same student and marker clusters with only one annotator will be filtered before the subsequent fitness value estimation process. Hence only meaningful markers were left for subsequent group formation.
- Finally, these triples will orchestrate the student characteristic matrix for genetic algorithm (Liang, Majumdar, Nakamizo, et al., 2022) with each marker cluster representing one row of the student matrix (*c*).

Throughout the workflow, we created context-oriented annotation indicators of active reading, thus enabling instructors to form groups with richer information about students in the active reading contexts, and laying the foundation of the empirical study.

4.2.3 Study Context and participants

To focus on the active reading activity, we choose a Japanese junior high school grade 2 English course as our research context. Three-day active reading-based group work activities were implemented in authentic English classes.

A total of 118 students (48 boys and 70 girls, with an average age of 14) from 3 classes instructed by the same English teacher participated in this study. There were 101 students (34 from Class A, 34 from Class B, and 33 from Class C) who actually participated in all three-day classes with others in their absence owing to personal issues. The consent form was first approved by the school authorities and then distributed to their parents/guardians as the students were minors. The parents signed the consent

form after being informed about the privacy issues of personal data collection and its usage for research only. We received consent from all 118 participating students.

4.2.4 Research design

Using a between-subjects design, we compared three conditions of group formation strategies in this study: heterogeneous groups with markers, heterogeneous groups with reading engagement, and baseline with random grouping. We confirmed that students in three classes who attended the study had no significant differences in their academic performance levels as scores in the latest mid-term English exam ($F = 0.553$, $p = 0.577$ for ANOVA test, see Table 4.5). Three different three group formation approaches were adopted: Class A using marker attributes (marker-based groups, MB), Class B using reading engagement attributes (engagement-based groups, EB), and Class C without attributes as a baseline condition (random groups). Students and the teacher did not know which condition each class belonged to throughout the experiment.

Table 4.5: ANOVA test for the scores of the latest exam of three classes

Class	Mean	SD	N	F	<i>p</i>
A	76.500	11.735	34	0.553	0.577
B	78.471	13.046	34		
C	75.424	11.264	33		

4.2.5 Procedure

The current study adopted a flipped learning activity, which lasted three days. Students first read the article "the history of clocks" proactively, thus creating the data for group formation. In this phase, the teacher instructed them to annotate the new words with yellow markers and important contents with red markers. Then they worked in groups to discuss the article and prepare for a group presentation. In the group presentation, students would perform a story based on their understanding of the article throughout the individual reading and group discussion phases. These group learning activities aimed to help students acquire new vocabulary and deepen their understanding of the article. Figure 4.9 shows the detailed process of the three-day experiment.

On Day 1, Students finished a pre-test of new vocabulary that would appear in the following reading material before the reading task. Then, they read the article independently

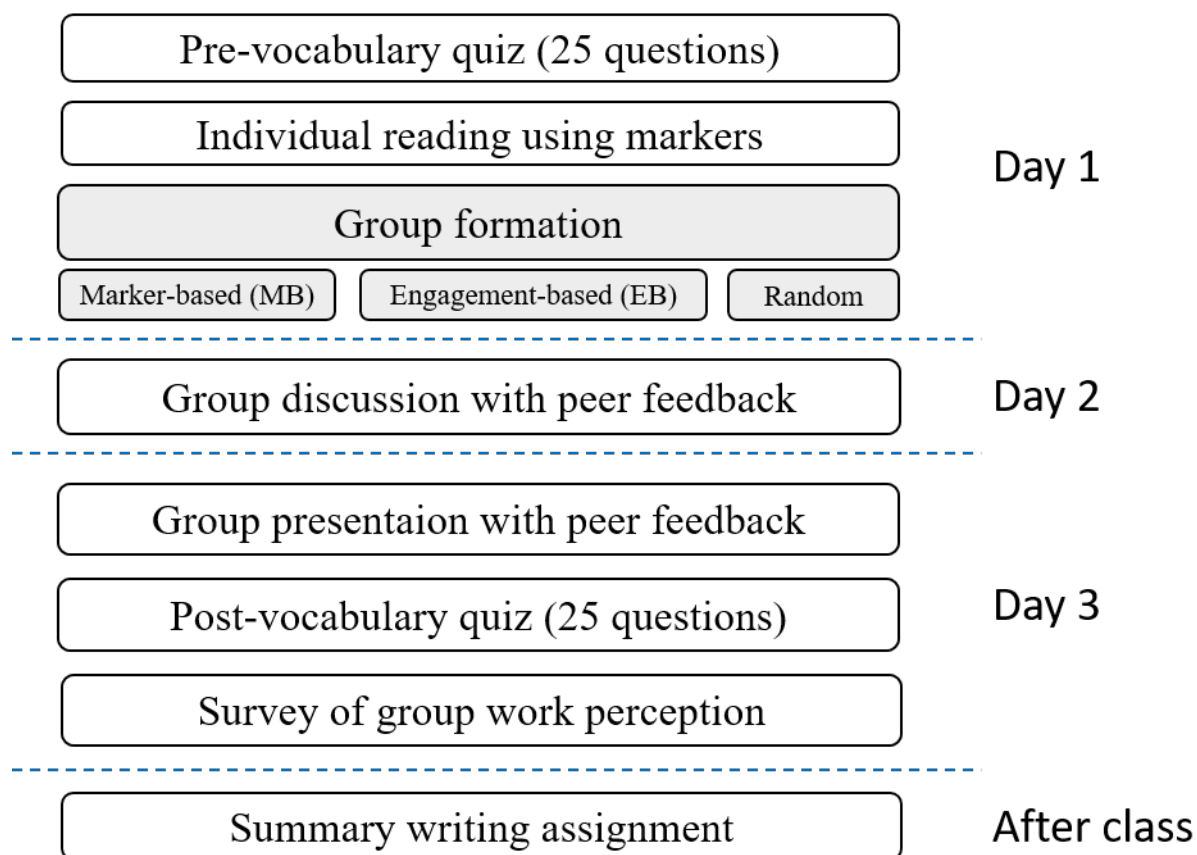


Figure 4.9: Procedure of the study

on the digital textbook BookRoll. Students were encouraged to use different markers and memos while reading according to the SQ4R strategy. In this case, the yellow marker annotated highlighted unknown words and expressions, and the red marker referred to important points and main ideas. Then, the teacher assigned students into groups (3 or 4 students per group) formed by the group formation system based on different conditions of reading data selection.

On Day 2, students discussed the article read on Day 1 with group members and prepared for the presentation based on the content and their understanding of the article. Students gave feedback on the participation and contribution of their group mates after the discussion.

On Day 3, each group gave a presentation in English in front of the class in 3 minutes. The teacher and audience made group evaluations with feedback on each group's presentation. Finally, students finished the post-test of vocabulary and took a short survey about their perception of the group work experience. After class, they independently

wrote a summary of the article as an assignment.

4.2.6 Data collection

In this study, we collected the rubric-based gradings of the summary writing task and performance scores on the vocabulary recognition quizzes (pre- and post-test). The summary writing task focused on reading comprehension and language use, and the vocabulary recognition quizzes measured new word learning. Further, a survey provided the student's perceptions regarding their group work.

Firstly, to measure the language learning achievement of students, the assignments about the summary of the article were collected by the memo function of BookRoll and graded by the teacher. The teacher graded the summary using scores ranging from 1 to 5 according to three criteria: content, word choice, and grammar. Figure 4.10 shows the rubrics used by the teacher for the summary grading.

Level	Word choice	Grammar	Content
5	Appropriate vocabularies that disclose the key concepts of the story and accurate grammar		Accurate and contains enough amount of information (covers main theme)
4	It has a few errors, but basically acceptable		Accurate but with slightly insufficient amount of information
3	Restricted vocabularies or grammar has some errors but is established as a sentence and acceptable		Dissatisfied but understandable
2	Many grammatical and vocabulary errors		Missing important points
1	Inappropriate / not established as a sentence		I can't understand at all

Figure 4.10: Rubrics used by teacher to grade summary

Secondly, we used a vocabulary recognition quiz with 25 questions in pre-test and post-test (0.4 points per question) because of the significant role of vocabulary as an appropriate assessment to measure reading performance and progress addressed by Richards and Burns (2012). One of the authors with more than ten years of English teaching experience created the vocabulary tests. Then the test items were checked by the teacher who taught the class before the disclosure to students. Each question presented an English word, and the student needed to choose the correct meaning in Japanese from the four choices. The words in the vocabulary quiz came from the new vocabulary list provided in the textbook "New Horizon" and the words that are necessary to understand the story. 88% of questions in the post-test appeared in the pre-test. This vocabulary recognition quiz

scores focuses on recognition of the new vocabulary while word choice score highlights the usage of keywords that are important in the story they read.

Thirdly, in terms of group work perception, a 5-item survey (see Table 4.9) was selected and adapted from Drury et al. (2003) with a 7-point Likert-type scale from “strongly agree” to “strongly disagree” to estimate the experience of the whole group work. The first two items measure the satisfaction of group work components which indicates subjective feelings about the group work, and the last three items involve self-evaluation of group work engagement and performance that denote group work dynamics. Around two-thirds of students finished the survey with others absent due to personal issues.

4.2.7 Data analysis

To answer RQ1, we analyzed data of summary writing scores and vocabulary recognition quiz scores. In terms of the summary writing tasks, the submission rate differed among three conditions, and EB groups had more students who completed the assignment ($n = 25, 74\%$) compared to MB groups ($n = 17, 50\%$) and random groups ($n = 17, 51\%$). Though the submission of the summary assignment was compulsory, it was hard to supervise each student after class. Since the teacher graded these summary assignments in five ordinal levels and the sample size was unbalanced, we conducted non-parameter examinations on the score of the summary assignment based on the Kruskal-Wallis H test and executed Dunn’s Post Hoc Comparisons to see the difference between each condition.

As for vocabulary recognition quizzes, we first conducted ANCOVA analysis to inspect the impact of different group formations on post-test scores. The test of homogeneity of regression showed the feasibility of ANCOVA with $F = .564$ ($p = .571$). To visualize such differences, we also calculated learning gains (LG) with normalized measures ranging from 0 to 1 according to (4.1) (Rebolledo-Mendez et al., 2022) and presented the LG for the three groups in a box plot.

$$LG = \frac{Posttest - Pretest}{10 - Pretest} \quad (4.1)$$

Meanwhile, to further investigate the transition patterns among the three conditions, we divided the quiz scores into three levels to examine the transitions of layers using the iSAT tool (Majumdar & Iyer, 2016). As a visual analytics tool for cohort analysis, the iSAT tool can illustrate the changes in the distribution of the dependent variable in the pre- and post-intervention phases with histograms.

As for the survey of group work perception to answer RQ2, Cronbach’s alpha value of this study was 0.885, suggesting relatively high reliability of the scales. Since the survey measurement was in ordinal levels, we investigated the difference between each item in three conditions with the Kruskal-Wallis H test and Dunn’s Post Hoc Comparisons as well.

4.2.8 Composition of marker-based groups

Table 4.6: Details of marker-based heterogeneous groups

	Group	# of unique markers	# (%) of markers from unique person
Yellow markers (difficult words)	1	41	39 (95.1%)
	2	54	46 (85.2%)
	3	54	51 (94.4%)
	4	22	19 (86.4%)
	5	53	48 (90.6%)
	6	39	39 (100%)
	7	15	11 (73.3%)
	8	21	20 (95.2%)
	9	40	32 (80.0%)
	10	6	6 (100%)
Red markers (highlights)	1	23	23 (100%)
	2	29	28 (96.6%)
	3	33	33 (100%)
	4	40	40 (100%)
	5	25	23 (92.0%)
	6	33	33 (100%)
	7	8	8 (100%)
	8	18	16 (88.9%)
	9	2	2 (100%)
	10	0	0

To check the performance of the updated algorithm using marker data. We examine the details of the group formation of the marker-based condition further. There were 102 meaningful yellow markers and 64 red markers annotated by at least two readers created on Day 1 for group creation. As shown in table 4.6, generally students with unique markers were evenly allocated into each group. More than 70% of markers annotated by the group members were different in each group for yellow markers, and the percentage was more than 85% for red markers. The results proved the reliability of the strategy of heterogeneously grouping students using annotation data.

4.2.9 Results

Performance in the summary assignment

Table 4.7 and 4.8 show that students in MB groups had a significantly higher performance in the word choice ($p = .003 < .01$) and grammar of their summary assignment ($p = .007 < .01$) than EB groups and random groups. However, the difference in these scores between EB and random groups was insignificant according to the post hoc test. Besides, we did not find the difference in the content score of the summary ($p = .774$), and the random groups had a higher mean score in this construct.

Table 4.7: Kruskal-Wallis test of the performance scores of the summary assignment

	Condition	N	Mean (SD)	H
Word choice	MB	17	4.588 (0.507)	11.657**
	EB	25	3.960 (0.889)	
	Random	17	3.824 (0.529)	
Grammar	MB	17	4.529 (0.514)	9.939**
	EB	25	3.960 (0.889)	
	Random	17	3.824 (0.529)	
Content	MB	17	3.059 (1.249)	0.512
	EB	25	3.320 (1.314)	
	Random	17	3.412 (1.372)	

** $p < .01$

Table 4.8: Dunn's Post Hoc Comparisons - scores of the summary assignment

	Comparison	W_i	W_j	z
Word choice	MB - EB	40.412	27.780	2.577**
	MB - Random	40.412	22.853	3.283***
	EB - Random	27.780	22.853	1.005
Grammar	MB - EB	39.471	28.160	2.322*
	MB - Random	39.471	23.235	3.054**
	EB - Random	28.160	23.235	1.011

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

Vocabulary quiz performance

From the pre-test to the post-test of the vocabulary recognition quiz, students in all three conditions improved their scores throughout the active reading-based group learning activities. Paired sample t-test showed the average quiz score of all students enhanced

from 7.22 to 8.82 ($t = 14.64, p < .001$). Figure 4.11 shows the box plot of the LG values in three conditions. MB groups (Mean = 0.645, SD = 0.29) had the highest LG, while EB groups (Mean = 0.541, SD = 0.378) had less LG than random groups (Mean = 0.587, SD = 0.298). MB groups had the lowest deviation of LG among the three conditions, indicating an equal improvement for all students. However, the difference in post-test scores among the three conditions did not reach a significant level in ANCOVA ($F = 0.872, p = .292$).

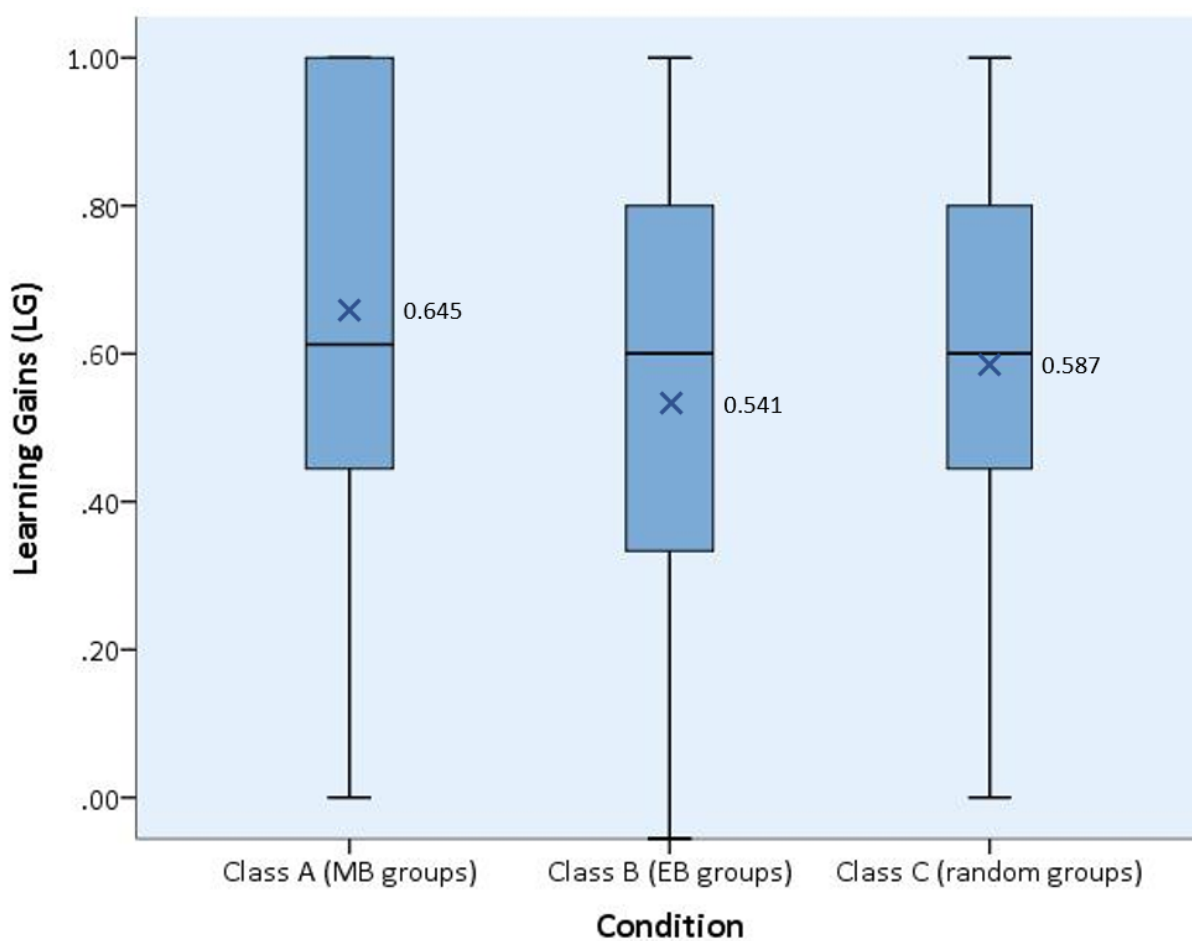


Figure 4.11: Box plot comparing learning gains in the vocabulary quizzes under three conditions

Figure 4.12 presents the transition patterns of the vocabulary quiz in three classes. Based on the iSAT diagrams, we can see that most students get 6 to 8 points in all three conditions in the pre-test. While in the post-test, all students in MB groups get more than 6 points, with 62% of them ($n = 21$) scoring more than 8 points. Meanwhile, those

who got more than 8 points in the pre-test ($n = 6$, 18%) remained the high score in the post-test. For EB groups, half of the students ($n = 17$, 50%) improved their score to more than 8 points, but there were two students whose scores decreased in the post-test. In random groups, two-thirds of the students ($n = 22$) got a post-test score higher than 8 points, while there was still one student who remained a low score below 6 points.

Survey of group work perception

Table 4.9 shows the statistics of each item from the group work perception survey. We can see students in MB groups tended to have higher scores in the equality of contribution (item 3) and self-evaluation of performance (items 4 and 5). However, only item 4 "I am a good player during the group work" showed statistical significance ($p = .013 < .05$). Post hoc comparisons showed a significant difference in the mean score of item 4 between MB groups (Mean = 6.095, SD = 0.831) and EB groups (Mean = 5.045, SD = 1.29) at the $p < .01$ level, and also between MB groups and random groups (Mean = 5.143, SD = 1.38) at the $p < .01$ level.

In items 1 and 2, the scores from the random grouping condition were even highest. The results indicated that the heterogeneous grouping strategy could not enhance the satisfaction and attitude towards the group work. Nevertheless, from items 3 to 5, we found that the marker-based group allocation enabled each individual to make a contribution to the group work and ensured the quality of the output. In addition, EB groups got the lowest average scores in all five survey items.

4.2.10 Discussion

Impact of marker-based grouping on language learning outcomes

As for English learning achievements throughout the experiment, we found a significant improvement in the vocabulary quiz scores for all three classes, which can support the related studies on group learning strategies in English class (Arisman & Haryanti, 2019; Muslim et al., 2022). This remarkable enhancement can be explained by the increased intrinsic motivations (Ehsan et al., 2019) and facilitated engagement in flipped learning scenarios (Acarol, 2019). Besides, the active reading strategy introduced on Day 1 can also contribute to such improvement according to Khusniyah and Lustyantje (2017). Beyond these studies on active reading and collaborative learning, this study investigated the impact of group learning based on reading annotation features to answer RQ1.

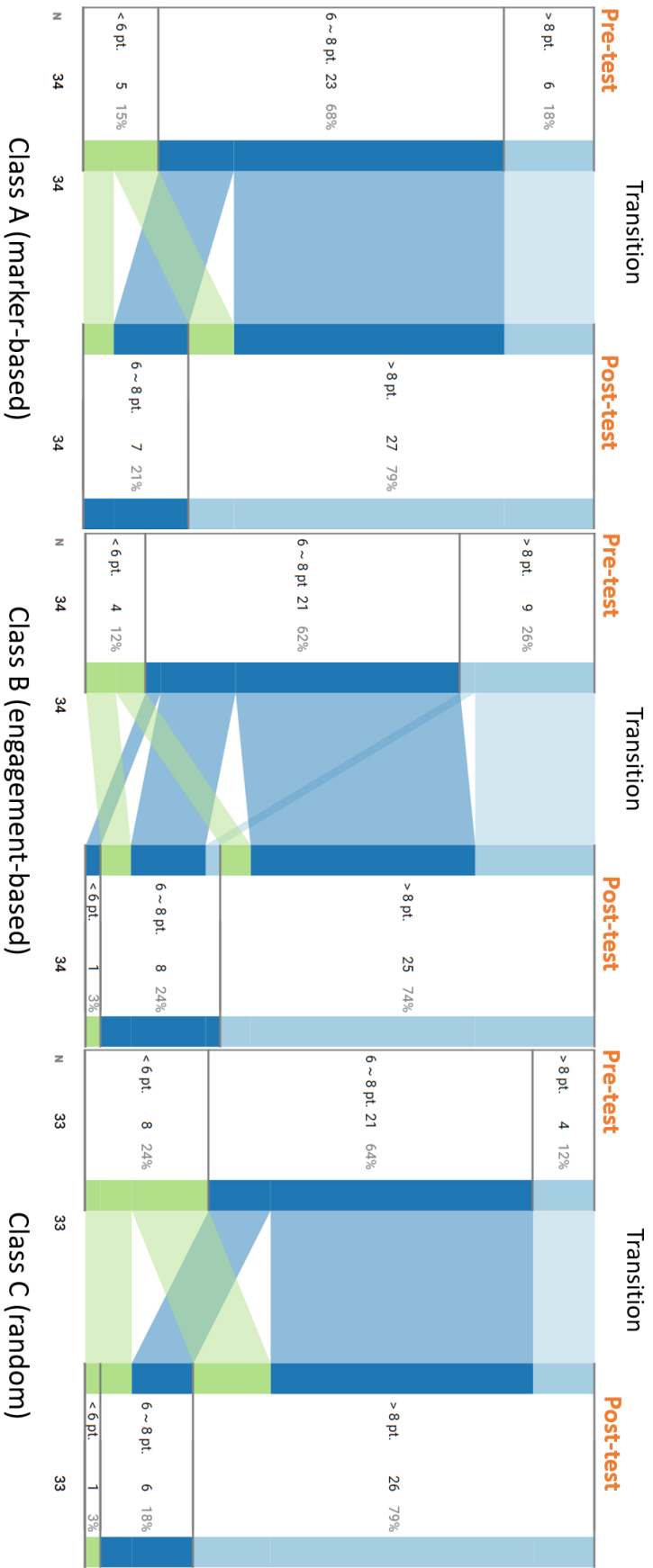


Figure 4.12: Transition graph of vocabulary quiz scores

Table 4.9: Survey items and Kruskal-Wallis test of responses

	Condition	N	Mean (SD)	H
1. I have had very positive experiences with group work.	MB	21	5.476 (1.078)	2.88
	EB	22	5.045 (1.588)	
	Random	21	5.762 (1.261)	
2. The product of group work has been as good or better than I could produce as an individual.	MB	21	5.476 (1.470)	3.329
	EB	22	5.091 (1.477)	
	Random	21	5.762 (1.480)	
3. We gave each member the opportunity to contribute.	MB	21	5.952 (1.071)	4.063
	EB	22	5.227 (1.343)	
	Random	21	5.714 (1.521)	
4. I am a good player during the group work.	MB	21	6.095 (0.831)	8.698*
	EB	22	5.045 (1.29)	
	Random	21	5.143 (1.38)	
5. We work well as a group.	MB	21	5.476 (1.25)	1.416
	EB	22	4.864 (1.642)	
	Random	21	5.000 (1.732)	

* $p < .05$.

As for the summary writing task, we can see better performance in the word choice and grammar scores in the condition with marker-based heterogeneous groups. To strengthen this finding, we inspected the pre-test score of the vocabulary recognition quiz for those who completed the assignment to control the effect of prior knowledge. Through ANOVA analysis, we found no significant difference in their pre-test scores ($F = 1.386$, $p = .259$), indicating their equal proficiency levels of vocabulary before the class. This finding can be explained by the knowledge level mentioned above as well. However, we could not see the difference in the content scores of the summary, which indicate the details of the article and should be meaningful for reading comprehension at the semantic level. Though we introduced the red markers that denote highlights of significant content as input of group formation, their impact on learning outcome was not detected. As was found by Sánchez et al. (2021), subjective characteristics such as personality traits and interests tend to have more impact on homogeneous groups. Therefore, the annotations of reading interest, which can vary from personal traits, can be more illustrative in homogeneous groupings. In the meantime, research on bibliographic coupling (Martyn, 1964) also inspired the potential for homogeneously created groups based on the common interest in active reading.

The results of the post-test score also denote the superiority of the marker-based

condition where students were allocated according to their annotations about difficult words and highlights. Such a heterogeneous strategy for peer help also agrees with related CSCL studies. According to the ZPD theory (Vygotsky, 1980), students with diverse annotations on difficult words formed an imbalance, thus laying a foundation to learn from each other. Existing studies can support our findings since the active reading annotations can be regarded as academic attributes as Han et al. (2020) proved, and can reflect the knowledge level according to Kanika et al. (2022). Compared to traditional group formation based on engagement indicators such as reading time, the marker attributes can reflect the knowledge level that is more predictive of the learning achievement of the group work. On the other hand, the engagement indicators can be more meaningful for individual learning (Chen et al., 2021) and self-direct learning (Li et al., 2021). As for group learning, though heterogeneous groups can guarantee that each group has at least one active student, problems such as neglect and isolation might happen (Salihoun et al., 2017).

In this study, we chose a heterogeneous algorithm for all input indicators due to the system limitation, which in turn led to further improvement of the group formation algorithm with a mixed grouping strategy. Nevertheless, despite the defects in highlight annotation handling for insignificant content score difference, the impact of active reading annotation data still exists on grammatical structure and vocabulary use of the summary writing assignment, which is explainable for active reading activity as well based on the position of vocabularies on reading assessment (Richards & Burns, 2012).

Effect of marker-based grouping on group work perception

According to the survey results for RQ2, subjective feelings about group work (items 1,2) and group work dynamics (items 3, 4, 5) are the studied constructs and we found no significant difference for them.

The satisfaction with the group work components indicated in items 1 and 2 is undesirable and even worse in the heterogeneous groups. This can be explained by Kanika et al. (2022), which admitted that despite the superiority of heterogeneous composition in academic achievement, students were more satisfied with homogeneous groups for subjective perception. Salihoun et al. (2017) also suggested homogeneous allocation in learning engagement patterns could reduce neglect and isolation of learners during group work, which may affect the subjective feelings of the participants.

Meanwhile, this study contributes to the group work practice in terms of group work engagement by using knowledge-level-based groups. Students in the marker-based groups gave higher scores than the other two conditions in all three items denoting the group work dynamics. This result agrees with the previous study that also investigated the engagement of knowledge level-based heterogeneous groups in a primary school context (Liang, Majumdar, & Ogata, 2021). However, the groups formed by heterogeneous reading engagement indicators tended to harm group work participation, which can be vulnerable to problems like social loafing and free-riding for low-engagement students (Strijbos, 2010).

Current Limitations

In this study, we inspected the effect of the marker-based group formation approach on language learning achievement and group work perceptions. When forming marker-based groups, we adopted a heterogeneous algorithm regardless of the marker type under the assumption that the unbalance of both vocabulary knowledge and reading interest within one group is meaningful in this learning context. However, its effect on learners' content comprehension remained unclear, since homogeneous red markers of readers that indicate common interest may enhance collaboration potential as well (Martyn, 1964; Sánchez et al., 2021). Since the red marker and yellow marker represent diverse active reading attributes that require different heterogeneity, there initiate adjustment of fitness function of the genetic algorithm (Liang, Majumdar, Nakamizo, et al., 2022) for mixed group formation that can accommodate both homogeneous and heterogeneous indicators (Revelo Sánchez et al., 2021).

As for input variables for group creation, this study focused on reading-related student model attributes only specific to different conditions to inspect our research questions in a language learning scenario. Nevertheless, smart learning platforms like LEAF (Ogata et al., 2022) make it possible to consider multiple variables for group formation for assorted learning contexts, which also broadens the horizons to weigh the impact of different student model attributes on the group work outcomes based on learning goals. For instance, in the subsequent group learning activities, teachers can group students with both annotation indicators and traditional engagement indicators for an overall consideration of their attributes.

For the empirical study, we faced obstacles when orchestrating group learning scenarios in the authentic junior high school context. The absence of students and the fluctuation

of the participation rate unbalanced the sample size of each measurement, which called for more rigorous instructions from teachers in future studies. Accordingly, only one item of the perception survey showed significance in the statistical test, making the evidence weak. Such limitations highlighted the importance of guaranteeing the completion of the measurement tasks. Besides, the knowledge level of the three actual classes may still vary despite their consistency in test scores. Though some studies on group learning adopted controlled experiments to avoid such problems, we stick to the empirical studies in a natural learning environment despite the trivial things to organize them since we aim to release real teachers from the difficulties of group work conduction with the smart systems. Meanwhile, the summary writing task was conducted after class due to the time limit of one class, and the unbalanced submission rate can make it harder to justify the out-performance of the marker-based group. As for the learning context, we shall think about the online grouping learning scenarios where virtual participation is allowed in future studies and try to overcome these drawbacks.

Implications for Technology and Pedagogy design

For technical contributions, we discoursed the data procession of reading annotation on e-book platforms to group students with similar or diverse markers. The study extended existing genetic group formation (Liang, Majumdar, Nakamizo, et al., 2022; Moreno et al., 2012) and suggests the potential to accommodate interactive features when calculating fitness values for optimal group compositions.

As for pedagogical implications, we applied the data-driven environment in classroom studies and proposed opportunities for group formation implementations using assorted learning log data, aiming to release teachers from difficulties when conducting group work using digital systems. This study showed the implications of reading annotations on difficult words (yellow markers), which could make high school teachers more competent in resembling classroom implementations. In agreement with Han et al. (2020), our findings demonstrated the feasibility of utilizing online learning data to obtain relevant information for effective face-to-face activity design in a flipped classroom. Meanwhile, we presented a general workflow to conduct classroom-based group learning activities supported by a data-driven environment under the GLOBE framework (Liang, Majumdar, & Ogata, 2021) and LEAF (Ogata et al., 2022) as is shown in Figure 4.13. The workflow is available not only in face-to-face language learning classes like this study but also in other reading-

based scenarios such as academic reading classes in higher-education level (Majumdar, Bakilapadavu, et al., 2021) and online courses (Majumdar, Flanagan, et al., 2021).

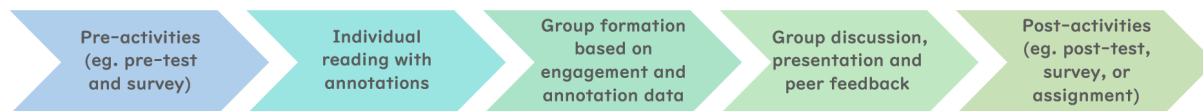


Figure 4.13: Example procedures for group work implementation under GLOBE framework and LEAF.

The study also presented a flipped-learning design. Following the procedure in Figure 4.13, students can proactively learn individually under the teacher’s instructions and the scaffold of the digital learning environment. They can work with recommended partners to prepare for the flipped presentation. During the presentation, they can reflect on their performance by exchanging feedback in the peer evaluation system as a formative assessment (Forsell et al., 2020). As for teachers, though they still need to prepare the quizzes and make appropriate instructions in each phase of the class, the group formation module can relieve them from the trivial work of grouping students and save time (Liang, Majumdar, & Ogata, 2021). In parallel, according to the limitation of this study, the measurement of the learning outcome of the flipped classroom also deserves our attention. To examine the effect of the flipped learning activity, teachers must give more rigorous instructions when measuring the learning achievement and leave enough time for the post-activities of the flipped classroom.

4.2.11 Conclusion and future work

In summary, this study connected two educational fields of language learning and CSCL by using active reading attributes to support collaborative learning by strengthening a data-driven group formation system with annotation data. Results from an empirical study found that students from the marker-based heterogeneous groups performed better in the vocabulary recognition quiz and the after-class summary writing assignment. The self-perception of their group work engagements was also higher in these groups. These findings supported the superiority of our innovative group formation technique. Through this study, we suggested opportunities for further group formation implementations using various data in the data-driven environment.

In future work, it is imperative to investigate the impact of highlight markers using the homogeneous composition as the limitation part mentioned, which can be significant

beyond language learning contexts. Accordingly, the group formation system needs further improvement to support a mixed group formation strategy. Meanwhile, the overall consideration of both annotation and engagement also deserves subsequent implementations.

Chapter 5

Utilize: Using group evaluation for subsequent group work

The GLOBE infrastructure offers more than just the consolidation of data from existing platforms. It also enables the creation and collection of data during group work, which can be utilized in subsequent rounds of the process. In this continuous data-driven workflow, peer evaluation modules play a pivotal role in gathering data on group work performance. This chapter presents a study that focuses on the cyclical utilization of peer evaluation data for algorithmic group formation.

5.1 Study 3: Group formation using continuously accumulated peer rating data

5.1.1 Aim and research question

In contrast to online learning environments, students in traditional classrooms have limited exposure to digital tools, resulting in a “cold start” problem due to the scarcity of learning logs required for creating learner models (Brusilovsky et al., 2015). Consequently, allocating students to groups based on their attributes becomes challenging. Addressing the gap highlighted in Figure 1.1, this study aims to explore the implementation and effectiveness of data-driven group formation, along with group work evaluation systems, within the context of a real junior high school classroom. 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 in 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?

5.1.2 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 understandings of the learning topic. Figure 5.1 shows the workflow of an actual idea exchange group learning (typical session 2 or 4).

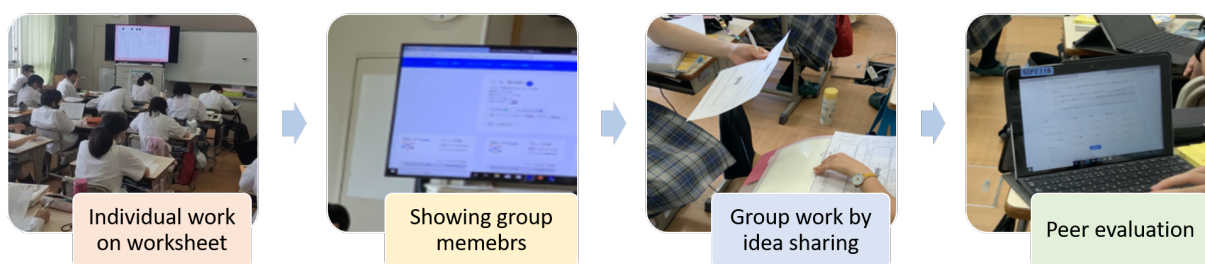


Figure 5.1: Idea exchange group learning: Classroom implementation workflow

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 students 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 5.2). 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.

	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			
Observation	Class B	Class A & B & C	Class B		Class A & C	
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 5.2: Procedure of the group learning experiment

5.1.3 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.

5.1.4 Procedure

The procedure of the study across four sessions was summarized in Figure 5.2. 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 the 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 sessions 2 and 3 were utilized following the continuous data flow in Figure 3.10.

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 & 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 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.

5.1.5 Instruments and data collection

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 (Liang, Toyokawa, et al., 2021). 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 5.1) was selected and adapted from the questionnaire of student perceptions 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 & 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 5.1: 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.

5.1.6 Data analysis

Before analysis, we conducted tests to confirm the equivalence of groups by considering their academic performance and attitude to group learning. Table 5.2 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

Table 5.2: 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		

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.

5.1.7 Results

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

Table 5.3 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 5.3 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.

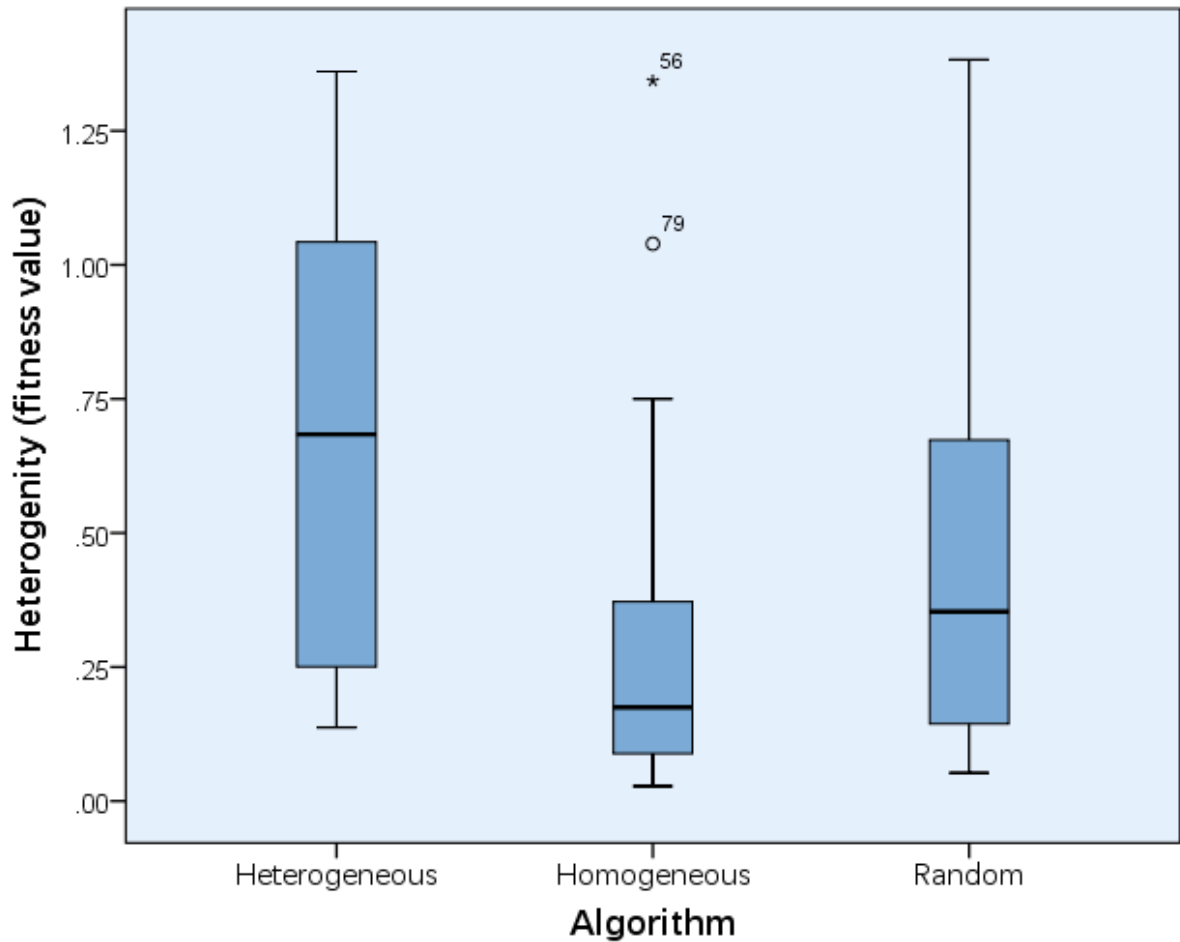


Figure 5.3: Box plot comparing heterogeneity of groups created by three approaches

Table 5.3: 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		

*** $p < .001$.

Table 5.4: 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**	.008
Homogeneous	Random	-0.107	-1.164	.478

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

Post-hoc tests found significant differences in heterogeneity between groups created by the 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 5.4).

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

Comparison of groups created by data-driven algorithmic group formation and random arrangement Table 5.5 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 5.5: 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

As is shown in Table 5.6 and 5.7, 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.

Table 5.6: Peer ratings of groups formed by the heterogeneous algorithm and random arrangement

	Group composition	N	Mean	SD	p	effect size
Subjectivity	Heterogeneous	39	3.876	0.976	.520	0.087
	Random	35	3.706	1.118		
Communication	Heterogeneous	39	3.769	1.012	.445	0.103
	Random	35	3.535	1.224		
Learning	Heterogeneous	39	3.829	0.983	.051	0.263
	Random	35	3.368	1.165		

As is shown in Table 5.8 and 5.9, 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 the homogeneous algorithm according to the self-perception survey of group work: the difference between two compositions on all peer

Table 5.7: Self-perception of group learning survey of groups formed by the heterogeneous algorithm and random arrangement

	Group composition	N	Mean	SD	<i>p</i>	effect size
Self-perception	Heterogeneous	34	18.206	4.176	.831	-0.031
	Random	32	18.176	4.421		

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.

Table 5.8: Peer ratings of groups formed by the homogeneous algorithm and random arrangement

	Group composition	N	Mean	SD	<i>p</i>	effect size
Subjectivity	Homogeneous	72	4.171	0.720	.003**	0.291
	Random	70	3.713	1.024		
Communication	Homogeneous	72	4.030	0.807	.037*	0.202
	Random	70	3.658	1.074		
Learning	Homogeneous	72	4.191	0.707	< .001***	0.354
	Random	70	3.677	0.963		

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

Table 5.9: Self-perception of group learning survey of groups formed by the homogeneous algorithm and random arrangement

	Group composition	N	Mean	SD	<i>p</i>	effect size
Self-perception	Homogeneous	68	19.324	2.878	.003**	0.305
	Random	58	17.483	3.521		

** $p < .01$.

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 (Fischer et al., 2002). We will inspect the results in two different contexts. Table 5.10 summarizes the comparisons under two different contexts.

Table 5.10: 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

As is indicated in Table 5.11 and 5.12, 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 perceptions 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.

Table 5.11: 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	.119	-0.178
	Homogeneous	72	4.171	0.720		
Communication	Heterogeneous	39	3.769	1.012	.097	-0.190
	Homogeneous	72	4.030	0.807		
Learning	Heterogeneous	39	3.829	0.983	.042*	-0.233
	Homogeneous	72	4.191	0.707		

* $p < .05$.

As is shown in Table 5.13 and 5.14, 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

Table 5.12: Self-perception of groups created by homogeneous and heterogeneous algorithms in idea exchange context

	Group composition	N	Mean	SD	<i>p</i>	effect size
Self-perception	Heterogeneous	34	18.206	4.176	.108	-0.194
	Homogeneous	68	19.324	2.878		

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.

Table 5.13: Peer ratings of groups created by homogeneous and heterogeneous algorithms in comparative reading context

	Group composition	N	Mean	SD	<i>p</i>	effect size
Subjectivity	Heterogeneous	70	3.735	0.862	.662	-0.051
	Homogeneous	37	3.712	1.022		
Communication	Heterogeneous	70	3.689	0.934	.293	-0.124
	Homogeneous	37	3.716	1.190		
Learning	Heterogeneous	70	3.783	0.731	.635	-0.056
	Homogeneous	37	3.676	1.155		

Table 5.14: Self-perception of groups created by homogeneous and heterogeneous algorithms in comparative reading context

	Group composition	N	Mean	SD	<i>p</i>	effect size
Self-perception	Heterogeneous	50	18.420	3.818	.297	0.141
	Homogeneous	29	17.138	5.370		

5.1.8 Discussion

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 in performance and outcome in the actual group work among groups with different heterogeneity values. According to the result of Figure 5.3, 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 delivers the flexibility that enables users to choose the algorithm as well as self-defined input variables, thus indicating potential implications in 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 values of 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 in 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 & Brown, 2006). For further improvement of the algorithm, more distance measures such as Cityblock, Euclidean, and Chebyshev should be considered as suggested by Flanagan et al. (2021).

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 (Liang, Majumdar, & Ogata, 2021). Figure 5.3 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 for 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 & 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 works 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 future studies.

Meanwhile, other factors may contribute to the observed small, non-significant differences between the homogeneous and heterogeneous groups such as the differences between Classes A, B, and C, and the sequencing of the sessions though we aimed to control them using test scores and group work attitude questionnaires. The imbalance of the samples in comparisons of homogeneous groups and heterogeneous groups could also affect the statistical results. As is shown in Table 5.12 and 5.13, 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 outcomes 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 & 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 weights 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 & Wasson, 2021), which would be considered in future system development.

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 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 & 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 M. Wang et al. (2020).

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 3.10 for the teacher in this study based on the learning context of a 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 (Liang, Majumdar, & Ogata, 2021) and English reading (Toyokawa et al., 2023).

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 in 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.

5.1.9 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 3.10).

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.

Chapter 6

Analyze: Recommendation of optimal group formation settings

With the accumulation of data under the GLOBE infrastructure, further data analysis can be conducted to help make decisions based on the insights gained from the previous two steps. This chapter will introduce a preliminary analysis of predictive group formation indicators for assisting teachers with automatic group formation.

6.1 Study 4: Predictive group work indicators for optimal group formation settings

6.1.1 Aim and research questions

To investigate the impact of each antecedent attribute, we run a correlation analysis using an online reading course under the LEAF and GLOBE infrastructure. The study aims to detect the relationship between the antecedent attributes and that in the subsequent phases (processes and consequences), which can be utilized to assist teachers to create groups with a recommendation of optimal group formation settings.

We conducted a single group study with a pulled-in dataset of one university course. During the weekly learning activities in the online learning platforms, their Collaborative Process Attributes were anonymously recorded in the data repository of GLOBE. This study aims to find optimal predictors for desirable group work by analyzing the correlation of the antecedents with processes and consequences attributes of collaboration. The overarching research questions of this study are as follows:

RQ1: What are the associations among individual-level indicators in different Collaborative Process Attributes?

RQ2: What are the associations among group-level indicators in different Collaborative Process Attributes?

6.1.2 Research context and participants

The dataset came from a university course "Readings in Humanities and Social Sciences: Education Technology and AI" in Japan in the academic year 2022. On completing this course, students should understand the structure and expressions in academic articles. The course also allowed students to improve their English reading and presentation skills. Weekly reading and group work activities were implemented under the LEAF and GLOBE infrastructure. The course collected abundant data on Collaborative Process Attributes, thus producing enough data samples from real-world settings with routine practices. Hence it holds generalizability (Maissenhaelter et al., 2018) and convenience for extraction of evidence in further analysis (Kuromiya et al., 2020). Thirty-two (32) students registered for the course at the beginning, with 7 students withdrawing midway. 25 students finished the whole course and got a final course grade. 19 students came from the Faculty of Engineering, 3 students came from the Faculty of Integrated Human Study, and the remaining 3 students majored in Pharmacy, Economics, and Science respectively. There were 17 sophomores, 5 junior students, and 3 senior students among the participants.

6.1.3 Procedure

In this course, group work was conducted several times from week 3 to week 11 across the 15-week semester. Following the GLOBE framework, students were grouped five times by the group formation system (Liang, Majumdar, & Ogata, 2021) across the course with different group formation indicators for different academic reading topics (see Table 6.1).

Figure 6.1 shows the workflow of the weekly activity implemented in the course. For each week, students were required to read several articles on BookRoll, an e-book reading system (Ogata et al., 2015) that can automatically collect learning data. Then, they should share and discuss their reading progress with their group members in the Moodle forum and prepare a brief presentation as a group for the next offline class. During the class, each group made presentations, which were peer-evaluated by the audience (both the instructor and students) in the classroom in the evaluation systems (Liang, Toyokawa,

Table 6.1: Group formation and group work topics in the course

	Input attributes	Group work topic	# of students	# of groups
Week 3-4	Reading engagement	Fast overview & reading strategy	32	6
Week 5-7	Reading engagement & previous peer ratings	Related work & review design	32	5
Week 8-9	Reading engagement	Keywords & systematic survey	25	5
Week 10	Reading engagement & previous peer ratings	Using group graphs	26	4
Week 11	Reading engagement & previous peer ratings	Using group graphs & self-directed learning	26	4

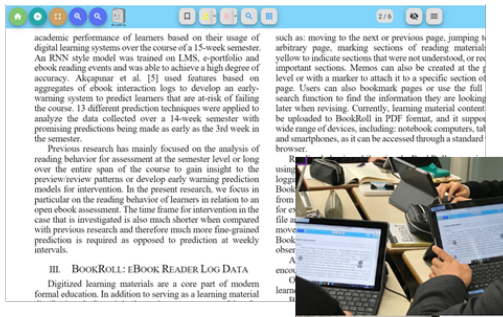
et al., 2021). In the meantime, students were asked to make peer ratings on the initiative and communication of their group mates in the peer evaluation system for each week as well.

6.1.4 Data collection

The data of 8 group work in 5 group compositions were pulled in for analysis since all of the group work followed the same procedure and identical rating rubrics. The individual indicators of antecedent attributes and process attributes were standardized into 0 to 1 for the group formation input. For group-level indicators, antecedent attributes were estimated by the squared differences (Flanagan et al., 2021) as heterogeneity and average scores were calculated for some process and consequence attributes (forum posts, forum characters, peer ratings of initiative, and peer ratings of communication). Table 6.2 summarizes all these indicators involved in the study.

6.1.5 Data analysis

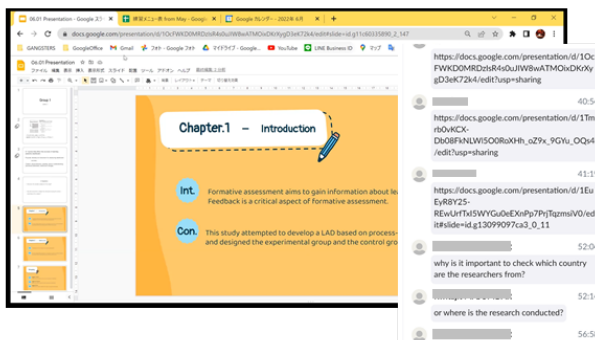
We used correlation analysis and calculated the Pearson correlation coefficient for each pair of antecedent-process and antecedent-consequence. To deal with missing values (eg. in weeks 3-4 and 8-9, previous group ratings and peer ratings as antecedent attributes



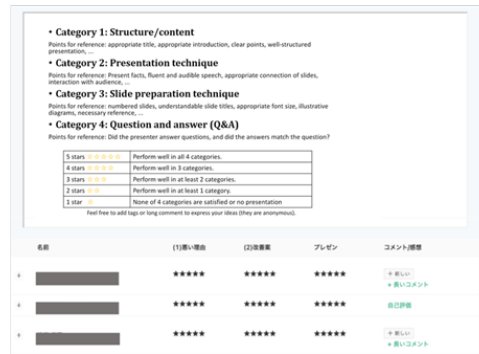
① Individual reading



② Forum discussion



③ Group presentation



④ Peer evaluation

Figure 6.1: Workflow of the weekly activity implemented in the course

were not used for group formation), we exclude cases pairwise before the analysis.

According to the research questions, we investigate two levels of indicators in this study. For individual-level indicators, we inspect the correlation among values. Positive relations denote that the higher score of an indicator one possesses, the more predictive of the desired learning outcome this indicator can be, and vice versa. Insignificant correlation means low predictive power in this learning context.

For group-level indicators, we examined their correlations with the group-level indicators of processes and consequences attributes that were calculated by aggregation of each group. The heterogeneity of each indicator as an antecedent attribute within a group is calculated by the squared differences, which are also used in the group formation algorithm to measure the heterogeneity of each group as the fitness function (Flanagan et al., 2021). As for the indicator of heterogeneity, the positive relation coefficient suggests the more heterogeneous the values of a certain indicator within a group, the better performance this group will have. On the contrary, negative correlations connote the more homogeneous the values of a certain indicator in a group, the more desirable the group-level outcome

Table 6.2: Indicators used in this study

Indicator	N	Mean	Max	Min
Antecedents				
Reading time	199	0.658	1	0.08
Operation times	199	0.653	1	0.06
Completion rate	199	0.483	0.65	0.05
Red markers	199	0.532	1	0
Yellow markers	199	0.551	1	0
Memos	199	0.384	1	0
*Heterogeneity of reading time	46	0.250	0.461	0.078
*Heterogeneity of operation times	46	0.239	0.409	0.035
*Heterogeneity of completion rate	46	0.123	0.218	0
*Heterogeneity of red markers	46	0.347	0.489	0.078
*Heterogeneity of yellow markers	46	0.335	0.526	0.064
*Heterogeneity of memos	46	0.388	0.509	0
Previous teacher's ratings	121	0.876	1	0.6
Previous peer ratings (individual)	109	0.738	1	0.2
Previous peer ratings (group)	121	0.786	0.9	0.629
*Heterogeneity of previous teacher's ratings	26	0.098	0.121	0.031
*Heterogeneity of previous peer ratings (individual)	26	0.243	0.312	0.076
*Heterogeneity of previous peer ratings (group)	26	0.091	0.132	0.017
Processes				
Forum posts	114	0.301	0.99	0
Forum characters	114	0.353	0.99	0
Consequences				
*Teacher's ratings	46	4.413	5	3
Peer ratings of initiative	199	3.658	5	0.5
Peer ratings of communication	199	3.461	5	1
*Peer ratings (group)	46	4.055	4.667	2.857
Final course grades	25	69.8	100	30

Note. * Group-level indicators.

will be.

6.1.6 Results

Individual-level indicators

Figure 6.2 is the correlation diagram of individual-level indicators. As can be seen in the diagram, reading time and previous peer ratings for individuals show significant positive associations to all processes and consequences attributes. The association between previous peer ratings for individual and final course grades is strong (> 0.7). Operation

times and the number of memos have significant positive correlations to all three consequence attributes, but their associations to process attributes are not found. Conversely, previous teachers' ratings of group work are related to the individual performance of two processes attributes, but not associated with all individual-level consequence scores. Both red markers and yellow markers take close relations to the final course grade. In addition, red markers show a weak significant association with initiative scores of peer ratings while yellow markers are weakly associated with communication scores of peer ratings. The completion rate connotes a weak adverse connection to the process attributes of forum utterance in this study and no significant correlation with all three consequence attributes. Meanwhile, previous peer ratings of group presentations indicate no significant relationship to any individual-level indicators.

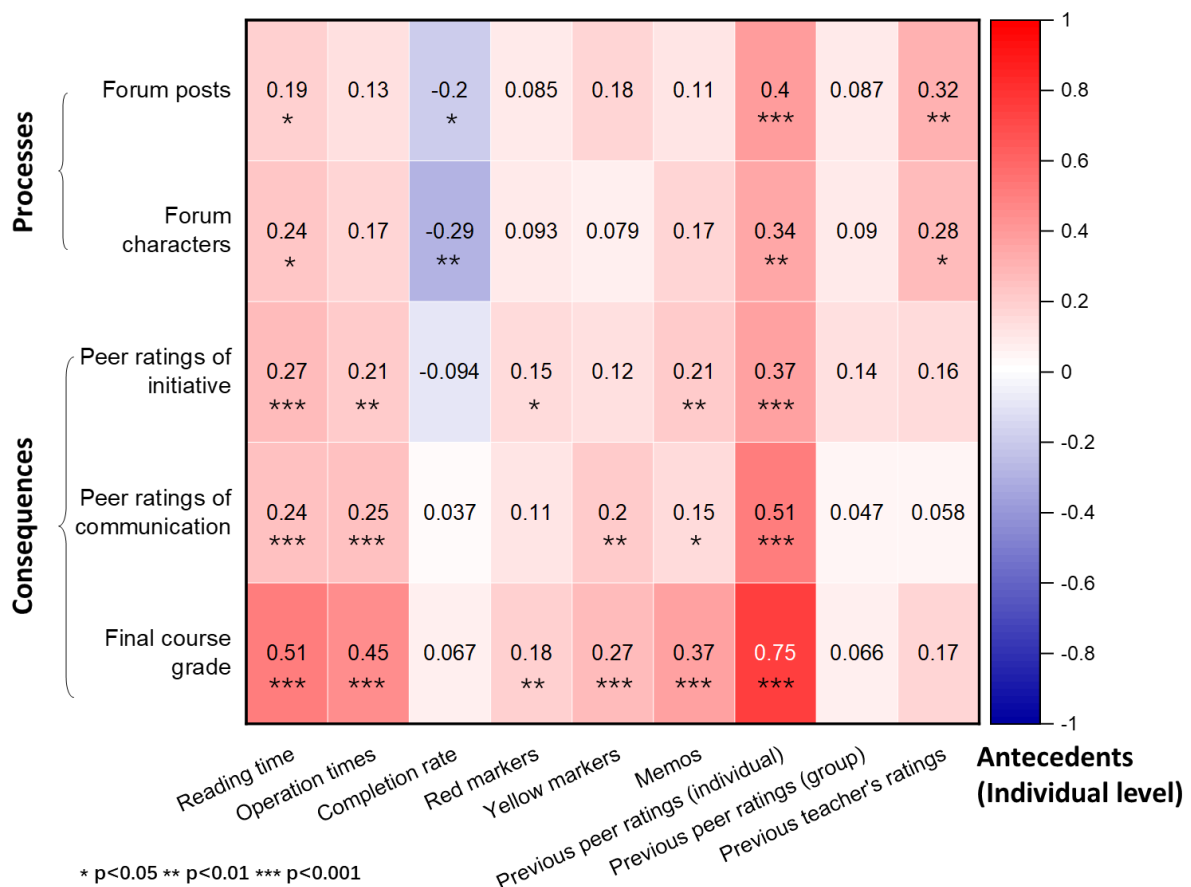


Figure 6.2: Results of correlation analysis of individual-level indicators of group work

Group level indicators

Figure 6.3 illustrates the results of correlation analysis for group-level indicators. As a result, positive and strong associations are found between (1) heterogeneity of previous peer ratings (group) and average forum posts and (2) heterogeneity of peer ratings (group) and average forum characters. This means that the heterogeneous composition of these antecedent attributes can contribute to the performance of the group work processes.

Negative correlations are revealed between (1) heterogeneity of red markers and average forum characters, (2) heterogeneity of previous peer ratings (individual) and the peer ratings received of the current group presentation, and (3) heterogeneity of previous peer ratings (group) and the peer ratings received of the current group presentation. These three correlations are moderate. This denotes the potential of the homogeneous composition on these antecedent attributes to scaffold the performance of the group work processes. Apart from the former results, all other correlations are insignificant in statistics.

6.1.7 Discussion

Individual-level indicators and individual performance

Compared to the previous study, most correlations in this study remain the same with Liang et al. (2022a). The reading time and previous peer ratings received are still the most predictive indicators that suggest a significant positive correlation with all processes attributes of forum engagement and consequences attributes of peer ratings as well as the final course grade. These results are also in accord with Junco et al. (2015) and Chen et al. (2021) that found reading time is predictive of the individual learning outcome. The active reading indicators such as memos and markers are also positively associated with desirable learning consequences as Yang et al. (2021) presented. In parallel, the reliability of peer ratings under the peer evaluation system can be approved as well, suggesting that students of the online university course can give a fair assessment to their peers based on rubrics. However, the completion rate showed an adverse association with the forum engagement indicators. This can be caused by the pull-in operation when we aggregate data. Since this study used all data from the course, the overall completion rate got lower due to the abundant reading materials as can be seen in the descriptive statistics in Table 3.

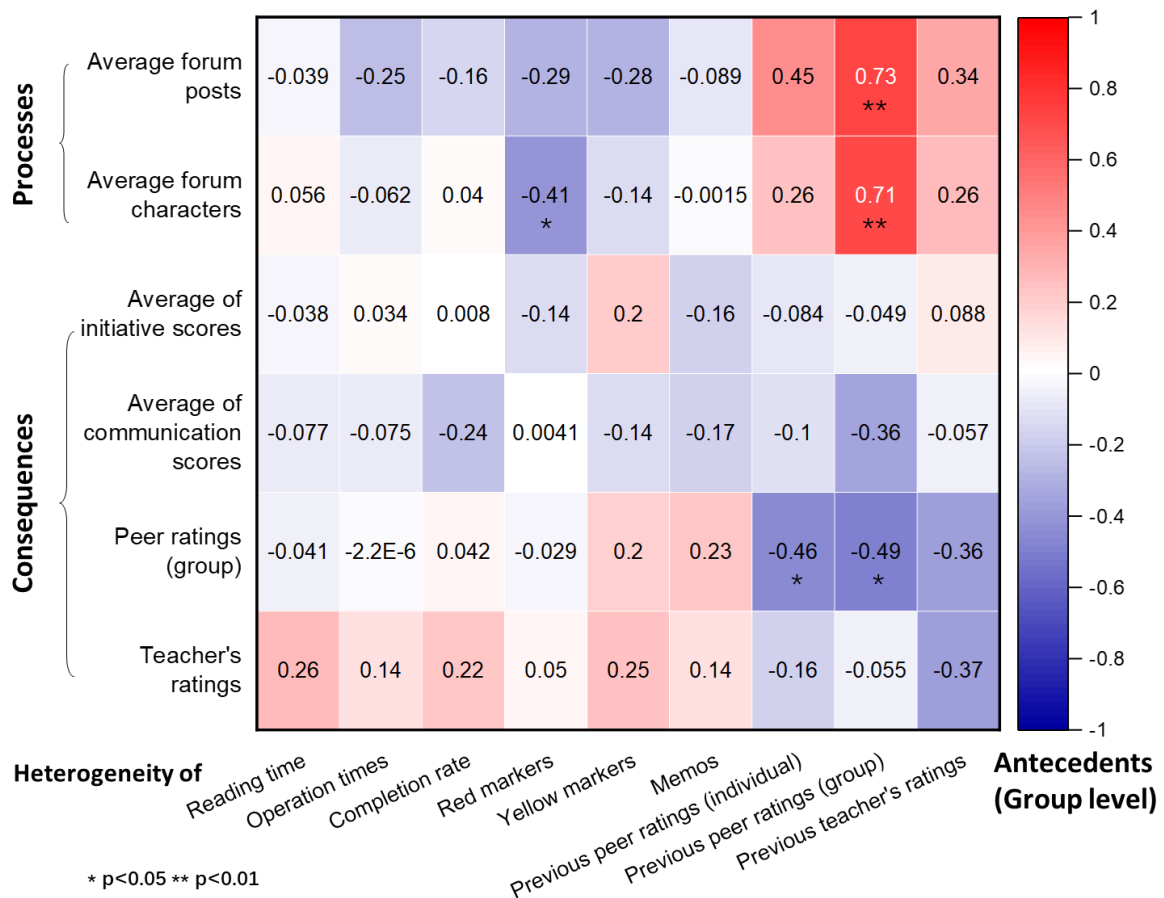


Figure 6.3: Results of correlation analysis of group-level indicators of group work

We can also find that, as for previous peer ratings and the teacher's ratings of group presentations, the predictive power is relatively low in that only the teacher's ratings of group presentations have a weak correlation to the forum engagement indicators. Since these two ratings are group-level assessments of previous group work, their reliability can be reduced by social loafing and free riding (Forsell et al., 2020), which can elicit less predictive power when modeling each individual using such scores. Apart from this, it also shows the necessity for analysis of group-level indicators as was mentioned by Cress (2008).

In sum, in the context of reading-based group work, the reading engagement attributes and peer ratings received in previous group work that indicates group work experience are closely connected to the individual performance of subsequent forum discussions and learning outcomes, which can guide group formation settings and intervention suggestions in the similar context such as language learning and academic reading. In parallel, from

the participation of reading and previous group work performance, teachers can take timely measures to help these endangered students predicted by the GLOBE system.

Group level indicators on group work performance

The group-level analysis focuses on the heterogeneity of each antecedent attribute within each group and aims to explore group dynamics. First, the average forum engagement of a group indicated by posts and characters is strongly positively correlated with its heterogeneity of previous group performance rated by peers. While no correlation was detected at the individual level between these two indicators. These findings support the strategy to heterogeneously group students so that we can guarantee that at least one outperforming student with desirable previous group work experience is assigned to each group, thus avoiding absolute silence in groups with all underperforming students. Such a positive effect of heterogeneous strategy on previous performance indicators agrees with group work in the classroom scenario as well (Liang, Majumdar, Nakamizo, et al., 2022).

As for annotation data that indicate the records of active reading strategy, we found the groups with more homogeneous red markers indicating highlights tend to have more forum discussions, though for individuals more markers did not indicate more posts. As an indicator of active reading engagement, the effect of grouping students with homogeneous engagement levels agrees with the other research on online courses and MOOCs (Abou-Khalil & Ogata, 2021; Sánchez et al., 2021), which can be explained by reduced social loafing for lack of proactive students to count on (Wichmann et al., 2016). Furthermore, the homogeneous grouping can be more promising when considering the annotated contents, since students with common annotations can show joint interest that can facilitate the interaction of the participants (Toyokawa et al., 2021).

Another finding that deserves our attention is that the heterogeneity of previous ratings, both for individuals and groups, are of moderate negative related to the peer rating scores of the final group presentation. The result denotes that though a group with heterogeneity in the previous group experience tends to have more discussion and engagement when it comes to the cooperative for a group-level output, it can become hard to reach a consensus, thus resulting in undesirable performance on group presentations. The heterogeneous groups with unbalanced knowledge of the task encourage peer help that facilitates individual achievement (Kanika et al., 2022), but it may not contribute to the cooperation and synergistic output of a group. To figure out the reason, further

analysis of forum discussions is required to investigate the relationship between processes and consequences indicators of group work in the orchestration phase of GLOBE.

According to our primary analysis (see Figure 6.4), we have identified appropriate group formation strategies for teacher assistance in the data-driven environment of LEAF. A homogeneous grouping strategy, considering the number of difficulty markers and previous peer ratings, has the potential to enhance the number of forum characters and peer ratings of group presentations. This finding provides guidance for subsequent group formation in the context of active reading-based group work. On the other hand, heterogeneous grouping based on previous peer ratings for groups can facilitate more detailed forum discussions with more characters in forum posts. This strategy can be useful for online courses where online reading and forum discussion are closely connected.

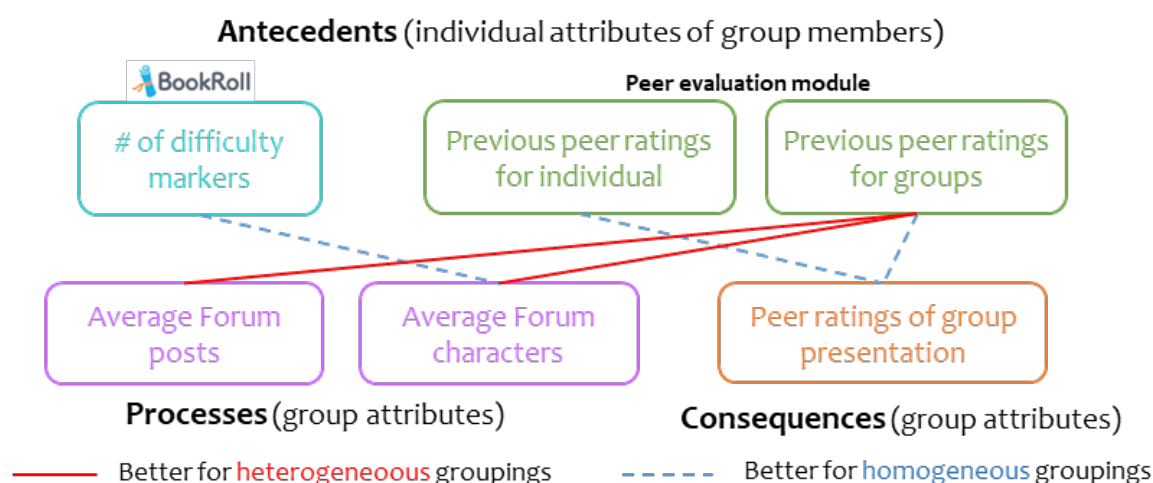


Figure 6.4: Suggested group formation strategies based on correlations between group-level attributes

Automatic group formation with optimal indicators to assist teachers

For technical implications, the research provides supportive evidence for the innovation of the current group formation system. Although we only addressed reading-based group discussions herein, similar research on other contexts can be done in the same way under the GLOBE framework. As is shown in Figure 6.5, teachers have to manually choose multiple indicators when creating groups currently. With the accumulation of evidence from studies on the predictive antecedent in different learning contexts, the strengthened system can automatically select input parameters based on the selected learning purpose

and context in the future. For example, homogeneous algorithms with red markers and previous peer ratings are suggested based on the results of this study, as they are associated with better group performance. Conversely, in contexts that underscore individual learning with peer help design, heterogeneous groups with reading engagement and test scores that indicate previous knowledge are recommended in the automatic grouping according to Liang, Majumdar, Nakamizo, et al. (2022) and Liang, Majumdar, and Ogata (2021).

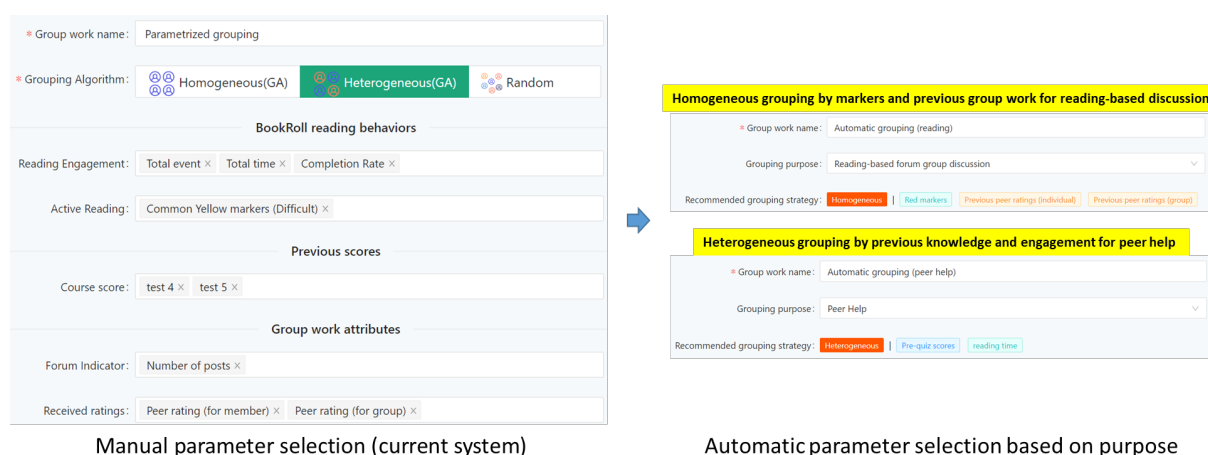


Figure 6.5: System innovation: From parameterized grouping to automatic grouping

For pedagogical implications, a pivotal goal of this study is to help teachers to determine the optimal group formation indicators in data-driven digital systems. This study discloses predictive antecedent indicators to the performance of subsequent group work in a forum-supported academic reading course, which can guide teachers in similar contexts. The automatic group formation function will further release teachers from selecting assorted variables in the system and reduce the time for creating groups. Further studies to examine the effectiveness of the automatic grouping will become necessary then.

Limitations

The indicators incorporated in this study are still limited. Under the data-driven platforms, most of the indicators are from learner models that reflect learning-related characteristics, but the social-emotional indicators are less addressed in the current systems. These issues should also be addressed by uploaded scores and social network data as quantitative input for group formation. However, how to incorporate these data with different

granularity and formats into the group formation algorithm remains unclear, and deserves future investigation. In parallel, the objective behavior data of previous group work was not used as the antecedent for the next round following the continuous data flow, which may reduce the reliability of previous group work performance indicators.

Meanwhile, though we got a larger sample size using pulled-in data of all group work throughout a semester in a university course compared to the previous study (Liang et al., 2022a), the learning context is confined to reading-based tasks with asynchronous forum discussions. Hence the predictive indicators in other learning conditions and cultures can vary. Therefore, the results of the current study need further validation in other learning scenarios.

6.1.8 Conclusion and future work

In conclusion, this study investigated the connections between antecedent attributes and the processes/consequences of group work in an asynchronous online reading course. We considered both individual-level and group-level indicators in the correlation analysis and found predictive indicators for algorithmic group formation. The reading engagement and previous peer ratings can reveal individual achievement of the group work, and a homogeneous grouping strategy based on reading annotations and previous group work experience can predict desirable group performance for this learning context. This study also provides avenues for future research to find predictive indicators in more learning contexts, and in turn, orchestrate an automatic group formation system that can mitigate teachers' trivial work from manual grouping. Meanwhile, how to make the antecedent indicators of groups created by algorithms explainable to teachers with adequate illustrations also deserves further consideration.

Chapter 7

Discussion

7.1 Summary of research

When conducting collaborative learning, grouping together learners with diverse strengths and weaknesses in the subject matter can provide individuals with opportunities to leverage their respective strengths. Conversely, forming groups comprising learners who share similar strengths and weaknesses allows them to concentrate on common challenging areas or enhance their proficiency in specific domains. However, the process of creating these groups has been arduous within everyday classroom settings due to time-consuming tasks like administering pre-tests and aggregating data. In this regard, learning analytics can play a crucial role.

The system discussed in this thesis facilitates group formation by utilizing learning log data. By leveraging this system, groups can be automatically generated using regular learning logs, thus lowering the barrier to incorporating group formation into everyday classroom activities. Remarkably, teachers have offered positive feedback, mentioning significant time savings in the group formation process, which has been streamlined from 1 to 1.5 hours down to approximately 30 minutes. Additionally, they have noted the system's ability to propose unexpected combinations that transcend conventional thinking.

Moreover, computer-based group formation offers distinct advantages over laborious manual processes. It provides a convenient platform for effortless manipulation, allowing for experimentation with diverse grouping conditions. This capability empowers educators to explore different approaches to forming groups based on their objectives. In this way, the group formation module of GLOBE, which utilizes learning logs, facilitates evidence-based and diverse group formation, thereby reducing the barrier to incorporating group

learning into everyday classroom activities.

In Figure 7.2, we present a data-driven perspective to address the three research questions. These studies encompass a range of learning contexts, spanning from primary school to higher education levels, wherein we applied the data-driven group formation system. A summary of these diverse contexts is depicted in Figure 7.1.





	Context	Learning logs	Grouping type	Task
Study 1	Mathematics Primary school Grade 5 (2019)	Grading scores Relationship	Heterogeneous Jigsaw re-grouping 	Problem solving
Study 2	English Middle school Grade 2 (2022)	Reading annotations Reading engagement	Heterogeneous 	Group discussions Group presentations
Study 3	Japanese Middle school Grade 2 (2021)	Peer ratings Teacher's ratings	Homogeneous  Heterogeneous 	Idea exchange & Comparative reading
Study 4	Academic reading University (2022)	Reading engagement Forum engagement Peer ratings		Forum discussion & Zoom presentations

Figure 7.1: Summary of research from the data-driven perspective.

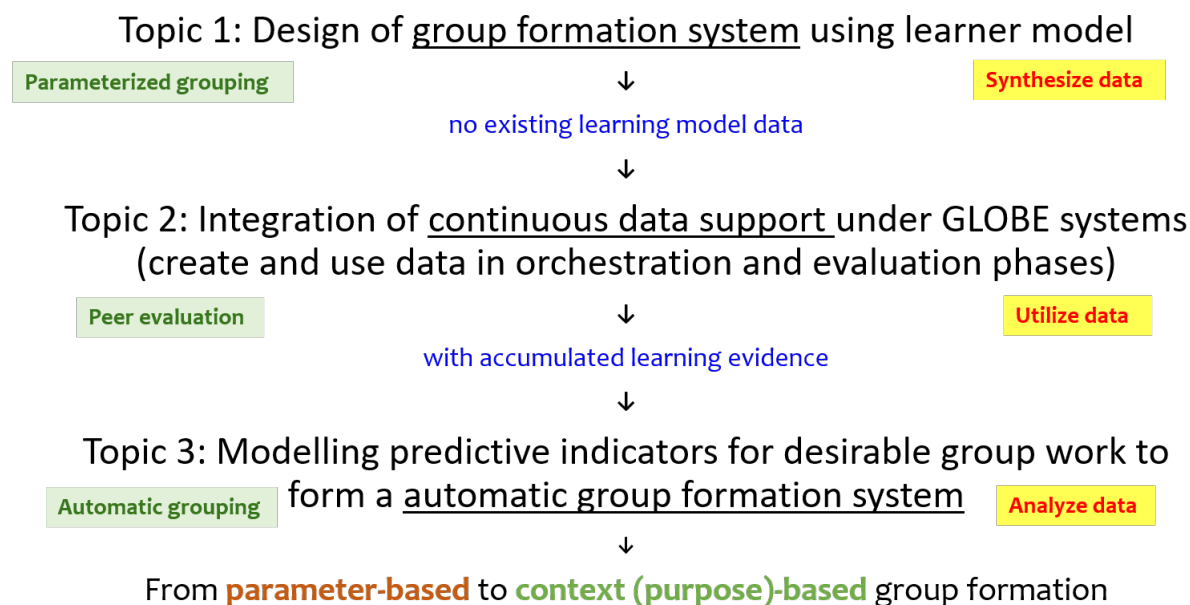


Figure 7.2: Summary of research from the data-driven perspective.

As for Topic 1, we initially implemented a group formation system with parameterized grouping to generate groups based on various learning log data, including course

scores, relationships, and shared annotations during active reading activities. The results demonstrated the system’s effectiveness in reducing time for teachers and enhancing group learning outcomes.

In situations where no pre-existing data is available for group formation, Topic 2 explored the potential utilization of a peer evaluation system to collect evidence of group learning, addressing the cold start issues commonly encountered in traditional classrooms. The findings indicated that data-driven groups created by the 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.

Building upon the accumulated evidence, data analysis was conducted to investigate predictive indicators of group formation in specific contexts. Topic 3 focused on a preliminary correlation analysis within a reading-based group learning environment. The results revealed that individual achievement in group work can be inferred from reading engagement and previous peer ratings. Moreover, a homogeneous grouping strategy based on reading annotations and prior group work experience can forecast favorable group performance in this particular learning context.

Based on these outcomes, our objective is to advance from parameter-based group formation to context-based group formation. In this approach, teachers only need to specify the purpose, and groups can be formed using optimized settings. By transitioning to context-based group formation, we aim to alleviate teachers from the trivial task of manually creating groups, empowering them to focus on other aspects of instruction.

7.2 Implications

From a technical perspective, our study contributes to the application and extensions of the general genetic algorithm in group learning, and the integration and reuse of group learning data as a lifecycle to support multiple rounds of group work and facilitate further research. We explore the use of multiple data sources in group formation based on Genetic Algorithms (GA) (Moreno et al., 2012), incorporating the concept of heterogeneity value derived from the fitness value within the genetic algorithm (Flanagan et al., 2021). Furthermore, we extend the genetic algorithm by considering relationship data (Study 1) and marker content overlaps (Study 2), allowing for adaptation to different learning contexts and explainable groups on mutual relationship. Meanwhile, our study goes beyond

a single episode and focuses on multiple rounds of group work by integrating and reusing group learning data throughout the cycle of GLOBE (Study 3). Drawing on continuous multiple group learning activity data, we provide a insight to explore the impact of group heterogeneity and specific characteristics on group work outcomes (Study 4). These data-driven approaches empower educators to facilitate and optimize group learning design, ultimately leading to improved educational outcomes.

The pedagogical implications of our study are twofold. Firstly, our group formation system offers teachers a streamlined approach to incorporating Computer-Supported Collaborative Learning (CSCL) into their instructional practices, relieving them from the laborious task of group formation. This allows teachers to allocate more time and energy to meaningful teaching and learning activities (Amarasinghe et al., 2021). Additionally, our research demonstrates the effectiveness of an iterative data flow that can adapt to situations where initial student model data is unavailable. This flexibility provides a low threshold for teachers to adopt data-driven group learning strategies, thereby promoting the use of CSCL in actual classroom activities. Moreover, the system leverages learning log data and algorithms to overcome the limitations of traditional grouping, which can be influenced by teachers' biases and unequal consideration of each student. Furthermore, we present example workflows of CSCL classroom design using the GLOBE system, showcasing the practical application of our research in various learning environments (see Figure 7.1). These examples provide valuable insights and opportunities for further application in broader learning contexts.

7.3 Limitations

First, the group formation algorithm employed in our study exhibits areas for improvement. To encompass a broader range of scenarios, it is advisable to consider the mixed group formation method that accommodates both homogeneous and heterogeneous indicators (Revelo Sánchez et al., 2021). Additionally, there is a need for further refinement of the input indicators used for group formation, with a focus on capturing more group dynamics rather than relying solely on subjective ratings. It is also recommended to incorporate a broader range of social-emotional indicators alongside learning analytics metrics. Future studies should also emphasize the minimization of differences among groups (Konert et al., 2014), a factor currently not fully considered in the existing system.

Regarding the continuous data flow in group learning, it is important to investigate the number of rounds of data accumulation required for the system to achieve optimal performance. This knowledge can contribute to the system’s acceptance and ease of use by teachers.

Furthermore, while this thesis primarily focuses on the group formation phase, it is essential to explore the other three phases of the GLOBE framework with equal attention, particularly the group learning process and inter-group interactions.

Finally, we must acknowledge that our empirical studies’ sample size and contextual scope have certain limitations. As we collected data from authentic classrooms in the real world, there were fluctuations in attendance due to personal circumstances. Consequently, the sample size may be restricted, potentially affecting the generalizability of the results. Additionally, our findings primarily stem from a specific learning context, namely reading-based group work. It remains unclear to what extent these findings can be generalized and adapted to other diverse learning contexts.

7.4 Future work

Regarding the group formation system, our future work involves the inclusion of additional indicators and the exploration of a mixed algorithm. we plan to investigate indicators derived from the orchestration phase, which provides insights into the dynamics of group work. These indicators can be derived from methods such as sequential analysis (Hoppe et al., 2021) and social network analysis (Saqr et al., 2020), offering a more comprehensive understanding of group interactions.

In addition to the focus on group formation, our research endeavors extend to data-driven support for the remaining three phases of the learning process. For instance, we are currently investigating methods to estimate the reliability of peer evaluations using learner model data before the activity (Liang, Gorham, et al., 2022). This estimation enables the identification of behaviors and unreliable raters during group learning activities involving peer evaluation. By calibrating peer ratings and alerting potential unserious raters beforehand, the integrity and accuracy of the peer evaluation process can be improved.

Chapter 8

Conclusion

In summary, this study focuses on leveraging data-driven approaches to enhance group learning support. To overcome challenges in group learning using learning analytics (LA), we introduce the Group Learning Orchestration Based on Evidence (GLOBE) framework, comprising a data-driven group formation system and peer evaluation module. Through a series of studies, we showcase the implementation of GLOBE in various contexts, demonstrating its positive impact on reducing the workload of teachers and enhancing group work outcomes. As we continue to accumulate more data within the GLOBE infrastructure, our ultimate goal is to establish an ecosystem for data-driven group learning, enabling teachers and students to benefit from data-driven insights and fostering effective group learning designs for the future.

Acknowledgement

Firstly I would like to sincerely express my gratitude to my supervisor, Prof. Hiroaki Ogata from *Kyoto University*. During the five years of research life, he usually shared his original ideas of the research which guide the direction of my research. He also spared no effort in sponsoring me to participate in several research projects and international conferences to get more experience in this research field. His encouragement inspired me to go through challenges and continue my study.

I am also grateful to my advisers: Prof. Mana Taguchi, Prof. Toshiyuki Shimizu, Prof. Donghui Lin, and Prof. Takayuki Ito for their insightful comments and critical assessment of every research progress report. Every time I give a progress report they patiently hear my report and gave me advice from their perspective without reservation. The discussion advisors not only provide me with the solution to many details of my research design but also broaden my horizons to related studies. Also, I am grateful to my thesis committee members, Professor Takayuki Ito and Professor Keishi Tajima, for their valuable comments that made this thesis better.

I would also like to thank lab members for their help to conduct my research and improve my thesis, especially mentors: Dr. Rwitajit Majumdar, Dr. Brendan Flanagan, Dr. Izumi Horikoshi, Dr. Yiling Dai from *Kyoto University*, Prof. Ulrich Hoppe from *University of Duisburg-Essen, Germany*, and Prof. Ivica Botički from *Faculty of Electrical Engineering and Computing, Zagreb, Croatia*, who gave me suggestions in my research career. I would like to thank the co-author students in my lab, primarily Ms. Yuko Toyokawa, Mr. Thomas Gorham, Mr. Yuta Nakamizo, and Mr. Taro Nakanish, who helped me a lot with the organization and consultation of empirical studies. Also, the support from the lab secretary and technical staff is indispensable, and the help from Ms. Noriko Nakajima over the past five years is exceptionally invaluable.

In addition, this endeavor would not have been possible without the generous support from the Japan Science and Technology Agency, which provides me with support funding

to allow me to dedicate myself to my research.

Finally, I would like to express my gratitude to my parents and family for their support and encouragement. With their support, I was able to continue my doctoral program and live without financial hardship. I would not be here without your sacrifices and love.

References

- Abnar, S., Orooji, F., & Taghiyareh, F. (2012). An evolutionary algorithm for forming mixed groups of learners in web based collaborative learning environments. *2012 IEEE international conference on technology enhanced education (ICTEE)*, 1–6.
- Abou-Khalil, V., & Ogata, H. (2021). Homogeneous student engagement: A strategy for group formation during online learning. *International Conference on Collaboration Technologies and Social Computing*, 85–92.
- Acarol, K. (2019). A study on the effectiveness of flipped learning model. *Kara Harp Okulu Bilim Dergisi*, *29*(2), 267–295.
- Amara, S., Macedo, J., Bendella, F., & Santos, A. (2016). Group formation in mobile computer supported collaborative learning contexts: A systematic literature review. *Journal of Educational Technology & Society*, *19*(2), 258–273.
- Amarasinghe, I., Hernández-Leo, D., & Ulrich Hoppe, H. (2021). Deconstructing orchestration load: Comparing teacher support through mirroring and guiding. *International Journal of Computer-Supported Collaborative Learning*, *16*(3), 307–338.
- Amason, A. C., & Sapienza, H. J. (1997). The effects of top management team size and interaction norms on cognitive and affective conflict. *Journal of management*, *23*(4), 495–516.
- Aminu, N., Hamdan, M., & 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.
- Andrade, H. G. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College teaching*, *53*(1), 27–31.
- Arisman, R., & Haryanti, I. S. (2019). Using small group discussion to improve students' reading achievement on narrative text. *English Community Journal*, *3*(1), 325–334.
- Aronson, E., & Bridgeman, D. (1979). Jigsaw groups and the desegregated classroom: In pursuit of common goals. *Personality and social psychology bulletin*, *5*(4), 438–446.
- Austin, R., Smyth, J., Rickard, A., Quirk-Bolt, N., & Metcalfe, N. (2010). Collaborative digital learning in schools: Teacher perceptions of purpose and effectiveness. *Technology, Pedagogy and Education*, *19*(3), 327–343.
- Ball, E., Franks, H., Jenkins, J., McGrath, M., & Leigh, J. (2009). Annotation is a valuable tool to enhance learning and assessment in student essays. *Nurse education today*, *29*(3), 284–291.
- Banihashem, S. K., Noroozi, O., van Ginkel, S., Macfadyen, L. P., & Biemans, H. J. (2022). A systematic review of the role of learning analytics in enhancing feedback practices in higher education. *Educational Research Review*, 100489.

- Benefield, G. A., Shen, C., & Leavitt, A. (2016). Virtual team networks: How group social capital affects team success in a massively multiplayer online game. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 679–690.
- Bjelobaba, G., Paunovic, M., Savic, A., Stefanovic, H., Doganjić, J., & Miladinovic Bogavac, Z. (2022). Blockchain technologies and digitalization in function of student work evaluation. *Sustainability*, 14(9), 5333.
- Boticki, I., Akçapınar, G., & Ogata, H. (2019). E-book user modelling through learning analytics: The case of learner engagement and reading styles. *Interactive Learning Environments*, 27(5-6), 754–765.
- Bremner, S. (2010). Collaborative writing: Bridging the gap between the textbook and the workplace. *English for Specific Purposes*.
- Brusilovsky, P., Somyürek, S., Guerra, J., Hosseini, R., Zadorozhny, V., & 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., & 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., & Andrews, B. (2002). Cognitive and psychological factors underlying secondary school students' feelings towards group work. *Educational Psychology*, 22(1), 75–91.
- Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325.
- Cen, L., Ruta, D., Powell, L., Hirsch, B., & Ng, J. (2016). Quantitative approach to collaborative learning: Performance prediction, individual assessment, and group composition. *International Journal of Computer-Supported Collaborative Learning*, 11, 187–225.
- Chang, M.-H., Kuo, R., Essalmi, F., Chang, M., Kumar, V., & Kung, H.-Y. (2017). Usability evaluation plan for online annotation and student clustering system—a tunisian university case. *Digital Human Modeling. Applications in Health, Safety, Ergonomics, and Risk Management: Ergonomics and Design: 8th International Conference, DHM 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part I 8*, 241–254.
- Chen, C. H., Yang, S. J., Weng, J. X., Ogata, H., & Su, C. Y. (2021). Predicting at-risk university students based on their e-book reading behaviours by using machine learning classifiers. *Australasian Journal of Educational Technology*, 37(4), 130–144.
- Cheng, W., & Warren, M. (1997). Having second thoughts: Student perceptions before and after a peer assessment exercise. *Studies in Higher Education*, 22(2), 233–239.
- Cleyen, O., Santa-Maria, G., Magdowski, M., & Thévenin, D. (2020). Peer-graded individualised student homework in a single-instructor undergraduate engineering course. *Research in Learning Technology*, 28.
- Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.) *Hillsdale, NJ: Lawrence Earlbaum Associates*.

- 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., & 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? *Collaborative-learning: Cognitive and computational approaches*. (pp. 1–19). Oxford: Elsevier.
- Dinh, J. V., Schweissing, E. J., Venkatesh, A., Traylor, A. M., Kilcullen, M. P., Perez, J. A., & Salas, E. (2021). The study of teamwork processes within the dynamic domains of healthcare: A systematic and taxonomic review. *Frontiers in Communication*, 6, 617928.
- Donnelly, P. J., Blanchard, N., Olney, A. M., Kelly, S., Nystrand, M., & D’Mello, S. K. (2017). Words matter: Automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 218–227.
- Draisbach, U., & Naumann, F. (2013). On choosing thresholds for duplicate detection. *Proceedings of the 18th International Conference on Information Quality (ICIQ)*.
- Drury, H., Kay, J., & Losberg, W. (2003). Student satisfaction with groupwork in undergraduate computer science: Do things get better? *Proceedings of the fifth Australasian conference on Computing education-Volume 20*, 77–85.
- Du, J., Fan, X., Xu, J., Wang, C., Sun, L., & Liu, F. (2019). Predictors for students’ self-efficacy in online collaborative groupwork. *Educational Technology Research and Development*, 67, 767–791.
- Ehsan, N., Vida, S., & Mehdi, N. (2019). The impact of cooperative learning on developing speaking ability and motivation toward learning english. *Journal of language and education*, 5(3 (19)), 83–101.
- Erkens, M., Manske, S., Hoppe, H. U., & Bodemer, D. (2019). Awareness of complementary knowledge in cscl: Impact on learners’ knowledge exchange in small groups. *International Conference on Collaboration and Technology*, 3–16.
- Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6), 304–317.
- Fidalgo-Blanco, Á., Sein-Echaluce, M. L., Garcíea-Peñalvo, F. J., & Conde, M. Á. (2015). Using learning analytics to improve teamwork assessment. *Computers in Human Behavior*, 47, 149–156.
- Fischer, F., Bruhn, J., Gräsel, C., & Mandl, H. (2002). Fostering collaborative knowledge construction with visualization tools. *Learning and Instruction*, 12(2), 213–232.
- Flanagan, B., Liang, C., Majumdar, R., & Ogata, H. (2021). Towards explainable group formation by knowledge map based genetic algorithm. *2021 International Conference on Advanced Learning Technologies (ICALT)*, 370–372.
- Forsell, J., Forslund Frykedal, K., & Hammar Chiriatic, E. (2020). Group work assessment: Assessing social skills at group level. *Small Group Research*, 51(1), 87–124.
- Fukazawa, M. (2010). Validity of peer assessment of speech performance. *ARELE: Annual Review of English Language Education in Japan*, 21, 181–190.

- Gillies, R. M. (2016). Cooperative learning: Review of research and practice. *Australian Journal of Teacher Education (Online)*, 41(3), 39–54.
- Gueldenzoph, L. E., & May, G. L. (2002). Collaborative peer evaluation: Best practices for group member assessments. *Business Communication Quarterly*, 65(1), 9–20.
- Han, J., Huh, S. Y., Cho, Y. H., Park, S., Choi, J., Suh, B., & Rhee, W. (2020). Utilizing online learning data to design face-to-face activities in a flipped classroom: A case study of heterogeneous group formation. *Educational Technology Research and Development*, 68(5), 2055–2071.
- Haq, I. U., Anwar, A., Rehman, I. U., Asif, W., Sobnath, D., Sherazi, H. H. R., & Nasralla, M. M. (2021). Dynamic group formation with intelligent tutor collaborative learning: A novel approach for next generation collaboration. *IEEE Access*, 9, 143406–143422.
- Harris, A. M., Gómez-Zarà, D., DeChurch, L. A., & Contractor, N. S. (2019). Joining together online: The trajectory of csw scholarship on group formation. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–27.
- Heck, T. (2013). Combining social information for academic networking. *Proceedings of the 2013 conference on Computer supported cooperative work*, 1387–1398.
- Hirashima, T., Yamasaki, K., Fukuda, H., & Funaoi, H. (2015). Framework of kit-build concept map for automatic diagnosis and its preliminary use. *Research and Practice in Technology Enhanced Learning*.
- Hoppe, H. U., Doberstein, D., & Hecking, T. (2021). Using sequence analysis to determine the well-functioning of small groups in large online courses. *International Journal of Artificial Intelligence in Education*, 31, 680–699.
- Hsu, T.-C., Abelson, H., Patton, E., Chen, S.-C., & Chang, H.-N. (2021). Self-efficacy and behavior patterns of learners using a real-time collaboration system developed for group programming. *International Journal of Computer-Supported Collaborative Learning*, 1–24.
- Huckman, R. S., Staats, B. R., & Upton, D. M. (2009). Team familiarity, role experience, and performance: Evidence from indian software services. *Management science*, 55(1), 85–100.
- Ifenthaler, D., & Yau, J. Y.-K. (2020). Utilising learning analytics to support study success in higher education: A systematic review. *Educational Technology Research and Development*, 68, 1961–1990.
- Isotani, S., Inaba, A., Ikeda, M., & Mizoguchi, R. (2009). An ontology engineering approach to the realization of theory-driven group formation. *International Journal of Computer-Supported Collaborative Learning*, 4, 445–478.
- Janssen, J., & Kirschner, P. A. (2020). Applying collaborative cognitive load theory to computer-supported collaborative learning: Towards a research agenda. *Educational Technology Research and Development*, 1–23.
- Jeong, H., Hmelo-Silver, C. E., & Jo, K. (2019). Ten years of computer-supported collaborative learning: A meta-analysis of cscl in stem education during 2005–2014. *Educational research review*, 28, 100284.
- Kanika, Chakraverty, S., Chakraborty, P., & Madan, M. (2022). Effect of different grouping arrangements on students' achievement and experience in collaborative learning environment. *Interactive Learning Environments*, 1–13.

- Kasch, J., van Rosmalen, P., Löhr, A., Klemke, R., Antonaci, A., & Kalz, M. (2021). Students' perceptions of the peer-feedback experience in moocs. *Distance Education*, 42(1), 145–163.
- Khusniyah, N. L., & Lustyantje, N. (2017). Improving english reading comprehension ability through survey, questions, read, record, recite, review strategy (sq4r). *English language teaching*, 10(12), 202–211.
- Kim, Y., D'Angelo, C., Cafaro, F., Ochoa, X., Espino, D., Kline, A., Hamilton, E., Lee, S., Butail, S., Liu, L., et al. (2020). Multimodal data analytics for assessing collaborative interactions. In M. Gresalfi & I. Horn (Eds.), *14th international conference of the learning sciences* (pp. 2547–2554). International Society of the Learning Sciences (ISLS).
- Kim, Y. J., Engel, D., Woolley, A. W., Lin, J. Y.-T., McArthur, N., & Malone, T. W. (2017). What makes a strong team? using collective intelligence to predict team performance in league of legends. *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 2316–2329.
- Konert, J., Bellhäuser, H., Röpke, R., Gallwas, E., & Zucik, A. (2016). Moodlepeers: Factors relevant in learning group formation for improved learning outcomes, satisfaction and commitment in e-learning scenarios using groupal. *Adaptive and Adaptable Learning: 11th European Conference on Technology Enhanced Learning, EC-TEL 2016, Lyon, France, September 13-16, 2016, Proceedings 11*, 390–396.
- Konert, J., Burlak, D., & Steinmetz, R. (2014). The group formation problem: An algorithmic approach to learning group formation. *Open Learning and Teaching in Educational Communities: 9th European Conference on Technology Enhanced Learning, EC-TEL 2014, Graz, Austria, September 16-19, 2014, Proceedings 9*, 221–234.
- Krouska, A., Troussas, C., & Sgouropoulou, C. (2023). A novel group recommender system for domain-independent decision support customizing a grouping genetic algorithm. *User Modeling and User-Adapted Interaction*, 1–28.
- Kudo, T. (2016). Tinysegmenter:javascript : "dake de kaka reta konpakutona wakachigaki sofutō ea".
- Kuromiya, H., Majumdar, R., & Ogata, H. (2020). Fostering evidence-based education with learning analytics. *Educational Technology & Society*, 23(4), 14–29.
- Kyndt, E., Raes, E., Lismont, B., Timmers, F., Cascallar, E., & 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.
- Li, H., Majumdar, R., Chen, M. R. A., & Ogata, H. (2021). Goal-oriented active learning (goal) system to promote reading engagement, self-directed learning behavior, and motivation in extensive reading. *Computers & Education*, 171, 104239.
- Liang, C., Gorham, T., Horikoshi, I., Majumdar, R., & Ogata, H. (2022). Estimating peer evaluation potential by utilizing learner model during group work. *Collaboration Technologies and Social Computing: 28th International Conference, CollobTech 2022, Santiago, Chile, November 8–11, 2022, Proceedings*, 287–294.
- Liang, C., Majumdar, R., Nakamizo, Y., Flanagan, B., & Ogata, H. (2022). Algorithmic group formation and group work evaluation in a learning analytics-enhanced environment: Implementation study in a japanese junior high school. *Interactive Learning Environments*, 1–24.

- Liang, C., Majumdar, R., & Ogata, H. (2021). Learning log-based automatic group formation: System design and classroom implementation study. *Research and Practice in Technology Enhanced Learning*, 16(1), 1–22.
- Liang, C., Toyokawa, Y., Nakanishi, T., Majumdar, R., & Ogata, H. (2021). Supporting peer evaluation in a data-driven group learning environment. *International Conference on Collaboration Technologies and Social Computing*, 93–100.
- Lin, H.-C., Hwang, G.-J., Chang, S.-C., & 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., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education*, 11(3), 279–290.
- Maissenhaelter, B. E., Woolmore, A. L., & Schlag, P. M. (2018). Real-world evidence research based on big data: Motivation—challenges—success factors. *Der Onkologe*, 24, 378–389.
- Majumdar, R., Akçapınar, A., Akçapınar, G., Flanagan, B., & Ogata, H. (2019). Laview: Learning analytics dashboard towards evidence-based education. *9th International Conference on Learning Analytics and Knowledge*, 386–387.
- Majumdar, R., Bakilapadavu, G., Majumder, R., Chen, M. R. A., Flanagan, B., & Ogata, H. (2021). Learning analytics of humanities course: Reader profiles in critical reading activity. *Research and Practice in Technology Enhanced Learning*, 16(1).
- Majumdar, R., Flanagan, B., & Ogata, H. (2021). Ebook technology facilitating university education during covid-19: Japanese experience. *Canadian Journal of Learning and Technology*, 47(4).
- Majumdar, R., & Iyer, S. (2016). Isat: A visual learning analytics tool for instructors. *Research and practice in technology enhanced learning*, 11(1), 1–22.
- Manske, S., Hecking, T., Chounta, I. A., Werneburg, S., & Ulrich Hoppe, H. (2015). Using differences to make a difference: A study on heterogeneity of learning groups. *O. Lindwall, P. Häkkinen, T. Koschmann, P. Tchounikine, & S. Ludvigsen (Eds.), Exploring the material conditions of learning: the computer supported collaborative learning (CSCL) conference 2015*, 1, 182–189.
- Manske, S., & Hoppe, H. U. (2016). The "Concept cloud": Supporting collaborative knowledge construction based on semantic extraction from learner-generated artefacts. *Proceedings - IEEE 16th International Conference on Advanced Learning Technologies, ICAALT 2016*.
- Maqtary, N., Mohsen, A., & Bechkoum, K. (2019). Group formation techniques in computer-supported collaborative learning: A systematic literature review. *Technology, Knowledge and Learning*, 24(2), 169–190.
- Martyn, J. (1964). Bibliographic coupling. *Journal of Documentation*, 20(4), 236.
- Masaki, U., Maomi, U. et al. (2008). Item response theory with assessors' parameters of peer assessment. *Journal of the Institute of Electronics, Information and Communication Engineers*, 91(2), 377–388.
- McDonald, D. W., & Ackerman, M. S. (2000). Expertise recommender: A flexible recommendation system and architecture. *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, 231–240.

- Mentzer, N., Laux, D., Zissimopoulos, A., & 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., & 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., & 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., & Wasson, B. (2021). Backward evaluation in peer assessment: A scoping review. *Computers & Education, 175*, 104319.
- Moreno, J., Ovalle, D. A., & 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., & 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.
- Muslim, M. I., Muslem, A., & Sari, D. F. (2022). Using small group discussion in teaching reading comprehension. *Research in English and Education Journal, 7*(1), 34–39.
- Nyikos, M., & Hashimoto, R. (1997). Constructivist theory applied to collaborative learning in teacher education: In search of zpd. *The modern language journal, 81*(4), 506–517.
- Ogata, H., Majumdar, R., AKÇAPINAR, G., HASNINE, M. N., & Flanagan, B. (2018). Beyond learning analytics: Framework for technology-enhanced evidence-based education and learning. *26th International Conference on Computers in Education Workshop Proceedings*, 493–496.
- Ogata, H., Majumdar, R., & Flanagan, B. (2023). Learning in the digital age: Power of shared learning logs to support sustainable educational practices. *IEICE TRANSACTIONS on Information and Systems, 106*(2), 101–109.
- Ogata, H., Majumdar, R., Yang, S. J., & Warriem, J. M. (2022). Learning and evidence analytics framework (leaf): Research and practice in international collaboration. *Information and Technology in Education and Learning, 2*(1), Inv–p001.
- Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-book-based learning analytics in university education. *International conference on computer in education (ICCE 2015)*, 401–406.
- Ollesch, L., Heimbuch, S., & Bodemer, D. (2019). Towards an integrated framework of group awareness support for collaborative learning in social media. *Proceedings of the 27th International Conference on Computers in Education*, 121–130.
- Picciano, A. G. (2012). The evolution of big data and learning analytics in American higher education. *Journal of asynchronous learning networks, 16*(3), 9–20.
- Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013). Tuned models of peer assessment in moocs. *Proceedings of The 6th International Conference on Educational Data Mining (EDM 2013)*.

- Pliakos, K., Joo, S.-H., Park, J. Y., Cornillie, F., Vens, C., & Van den Noortgate, W. (2019). Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. *Computers & Education, 137*, 91–103.
- Pöysä-Tarhonen, J., Care, E., Awwal, N., & Häkkinen, P. (2018). Pair interactions in online assessments of collaborative problem solving: case-based portraits. *Research and Practice in Technology Enhanced Learning*.
- Rebolledo-Mendez, G., Huerta-Pacheco, N. S., Baker, R. S., & du Boulay, B. (2022). Meta-affective behaviour within an intelligent tutoring system for mathematics. *International Journal of Artificial Intelligence in Education, 32*(1), 174–195.
- Rentsch, J. R., & Klimoski, R. J. (2001). Why do ‘great minds’ think alike?: Antecedents of team member schema agreement. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, 22*(2), 107–120.
- Revelo Sánchez, O., Collazos, C. A., & Redondo, M. A. (2021). Automatic group organization for collaborative learning applying genetic algorithm techniques and the big five model. *Mathematics, 9*(13), 1578.
- Richards, J. C., & Burns, A. (2012). *The cambridge guide to pedagogy and practice in second language teaching*. Cambridge University Press.
- Rodríguez-Triana, M. J., Martínez-Monés, A., Asensio-Pérez, J. I., & 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., & Suwignyo, H. (2021). Assessment of learning development to improve student’s appreciative and critical thinking abilities in drama appreciation course. *4th Sriwijaya University Learning and Education International Conference (SULE-IC 2020)*, 495–502.
- Salihoun, M., Guerouate, F., Berbiche, N., & 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., & 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., & Bote-Lorenzo, M. L. (2019). Creating collaborative groups in a mooc: A homogeneous engagement grouping approach. *Behaviour & Information Technology, 38*(11), 1107–1121.
- Saqr, M., Nouri, J., Vartiainen, H., & Malmberg, J. (2020). What makes an online problem-based group successful? a learning analytics study using social network analysis. *BMC medical education, 20*, 1–11.
- Savicki, V., Kelley, M., & Lingenfelter, D. (1996). Gender, group composition, and task type in small task groups using computer-mediated communication. *Computers in human behavior, 12*(4), 549–565.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3/4), 591–611.

- 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. *Proceedings of the 2nd international conference on learning analytics and knowledge*, 4–8.
- Sivaloganathan, S., Al-Marzouqi, A., & Zanelidin, E. (2020). Teaching conceptual design to a heterogeneous group: A workshop method. *2020 ASEE Virtual Annual Conference Content Access*.
- Slof, B., van Leeuwen, A., Janssen, J., & Kirschner, P. A. (2021). Mine, ours, and yours: Whose engagement and prior knowledge affects individual achievement from online collaborative learning? *Journal of Computer Assisted Learning*, 37(1), 39–50.
- Smith, J., Bratt, H., Richey, C., Bassiou, N., & Alozie, N. (2016). Spoken interaction modeling for automatic assessment of collaborative learning. *Speech Prosody 2016*.
- Stahl, G., Koschmann, T., & Suthers, D. D. (2006). Computer-supported collaborative learning: An historical perspective. *Cambridge handbook of the learning sciences*, 409–426.
- Strauß, S., & 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.
- Strode, D., Dingsøyr, T., & Lindsjorn, Y. (2022). A teamwork effectiveness model for agile software development. *Empirical Software Engineering*, 27(2), 56.
- 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. *7th International Conference on University Learning and Teaching (InCULT 2014) Proceedings*, 603–616.
- Toyokawa, Y., Majumdar, R., Kondo, T., Horikoshi, I., & Ogata, H. (2023). Active reading dashboard in a learning analytics enhanced language-learning environment: Effects on learning behavior and performance. *Journal of Computers in Education*, 1–28.
- Toyokawa, Y., Majumdar, R., Lecailliez, L., Liang, C., & Ogata, H. (2021). Technology enhanced jigsaw activity design for active reading in english. *2021 International Conference on Advanced Learning Technologies (ICALT)*, 367–369. <https://doi.org/10.1109/ICALT52272.2021.00118>
- Urhahne, D., Schanze, S., Bell, T., Mansfield, A., & 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., & 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., & 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, C., & Xu, Y. (2023). Who will work together? factors influencing autonomic group formation in an open learning environment. *Interactive Learning Environments*, 1–19.
- Wang, M., Guo, W., Le, H., & 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., & Pfister, H.-R. (2001). Group formation in computer-supported collaborative learning. *Proceedings of the 2001 international ACM SIGGROUP conference on supporting group work*, 24–31.
- Wichmann, A., Hecking, T., Elson, M., Christmann, N., Herrmann, T., & Hoppe, H. U. (2016). Group formation for small-group learning: Are heterogeneous groups more productive? *Proceedings of the 12th international symposium on open collaboration*, 1–4.
- Willey, K., & 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.
- Winkler, W. E., & Thibaudeau, Y. (1991). *An application of the fellegi-sunter model of record linkage to the 1990 us decennial census*. Citeseer.
- 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.
- Xu, L., Zhou, X., & Gadiraju, U. (2020). How does team composition affect knowledge gain of users in collaborative web search? *Proceedings of the 31st ACM conference on hypertext and social media*, 91–100.
- Zamani, M. (2016). Cooperative learning: Homogeneous and heterogeneous grouping of iranian efl learners in a writing context. *Cogent Education*, 3(1), 1149959.
- Zheng, Z., & Pinkwart, N. (2014). A discrete particle swarm optimization approach to compose heterogeneous learning groups. *2014 IEEE 14th international conference on advanced learning technologies*, 49–51.