Group optimization to improve peer assessment accuracy using item response theory and integer programming

Nguyen Duc Thien

Graduate School of Information Systems
The University of Electro-Communications

A dissertation submitted in partial satisfaction
of the requirements for the degree of
Doctor of Philosophy in Engineering

March 2018

Group optimization to improve peer assessment accuracy using item response theory and integer programming

Approved by the Supervisory Committee:

Professor Maomi Ueno
Chairman

Professor Akihiko Ohsuga

Professor Satoshi Kurihara

Professor Yasuhiro Minami

Associate Professor Shuichi Kawano

Date Approved by the Chairman: _____

Date of the Defense: 22nd January 2018

項目反応理論と整数計画法を用いたピアアセスメントの精度向上のためのグループ最適化

Nguyen Duc Thien

概要

近年，MOOCsなどの大規模型eラーニングが普及してきた．大規模な数の学習者が参加している場合には，教師が一人で学習者のレポートやプログラム課題などを評価することは難しい．大規模の学習者の評価手法の一つとして，学習者同士によるピアアセスメントが注目されている．MOOCsのように学習者数が多い場合のピアアセスメントは，評価の負担を軽減するために学習者を複数のグループに分割してグループ内のメンバ同士で行うことが多い．しかし，この場合，グループ構成の仕方によって評価結果が大きく変化してしまう問題がある．この問題を解決するために，本研究では，項目反応理論と整数計画法を用いて，グループで行うピアアセスメントの精度を最適化するグループ構成手法を提案する．具体的には，項目反応理論において学習者の能力測定精度を表すフィッシャー情報量を最大化する整数計画問題としてグループ構成問題を定式化する．実験の結果，ランダムグループ構成と比べて，提案手法はおおむね測定精度を改善したが，それは限定的な結果であることが明らかとなった．そこで，本研究ではさらに，異なるグループから数名の学習者を外部評価者として各学習者に割り当てる外部評価者選択手法を提案する．シミュレーションと実データ実験により，提案手法を用いることで能力測定精度を大幅に改善できることを示す．

Group optimization to improve peer assessment accuracy using item response theory and integer programming

Nguyen Duc Thien

Abstract

In recent years, large-scale e-learning environments such as Massive Online Open Courses (MOOCs) have become increasingly popular. In such environments, peer assessment, which is mutual assessment among learners, has been used to evaluate reports and programming assignments. When the number of learners increases as in MOOCs, peer assessment is often conducted by dividing learners into multiple groups to reduce the learners' assessment workload. In this case, however, the accuracy of peer assessment depends on the way to form groups.

To solve the problem, this study proposes a group optimization method based on item response theory (IRT) and integer programming. The proposed group optimization method is formulated as an integer programming problem that maximizes the Fisher information, which is a widely used index of ability assessment accuracy in IRT. Experimental results, however, show that the proposed method cannot sufficiently improve the accuracy compared to the random group formulation.

To overcome this limitation, this study introduces the concept of external raters and proposes an external rater selection method that assigns a few appropriate external raters to each learner after the groups were formed using the proposed group optimization method. In this study, an external rater is defined as a peer-rater who belongs to different groups. The proposed external rater selection method is formulated as an integer programming problem that maximizes the lower bound of the Fisher information of the estimated ability of the learners by the external raters. Experimental results using both simulated and real-world peer assessment data show that the introduction of external raters is useful to improve the accuracy sufficiently. The result also demonstrates that the proposed external rater selection method based on IRT models can significantly improve the accuracy of ability assessment than the random selection.

Acknowledgements

I would like to express my appreciation to the people who have always encouraged and supported me in graduate studies at the University of Electro-Communications (UEC).

Foremost, I would like to express my sincere gratitude to my supervisor, Professor Maomi Ueno, for all his enthusiastic support and guidance in the past six years. Without his patience, encouragement, and guidance, this work would not have been possible. I would like to gratefully thank the dissertation committee, Professor Akihiko Ohsuga, Professor Satoshi Kurihara, Professor Yasuhiro Minami, and Associate Professor Shuichi Kawano, for their time in serving as the committee members and for their insightful comments and suggestions. I would like to particularly thank Assistant Professor Masaki Uto, who has been my mentor for the past three years. I am grateful for his enthusiastic guidance, support, and collaboration during my Ph.D. studies.

I would like to thank Professor Yutaka Ikeda, Professor Yuko Takeda, and Associate Professor Tetsuko Hamano of the Center for International Programs and Exchange (CIPE), UEC for their support in improving my Japanese and in finding scholarships for my studies. I would also like to express my appreciation to the officers of the International Student Office for their kindness during my student life at UEC.

I would like to acknowledge the financial support from the Tatsunoko Foundation for the past five years. I am also grateful to the Chairman Tatsuya Akimoto, the Managing Director Yuichi Shiitsuka, and Ms. Yukiko Kato, for their warm support and encouragement to me as a fellow of the Tatsunoko fellowship. I would like to express my gratitude to the University of Electro-Communications for the financial aid and for providing me opportunities to pursue my studies. I also gratefully acknowledge the research funding from the JSPS KAKENHI grants.

I thank my colleagues at the Ueno and Kawano laboratory, my fellows of the UEC Aikido club, my Vietnamese friends at UEC, and all other friends, who kindly helped me and shared with me the moments that made my student life in Japan memorable.

Finally, I would like to thank my family for all their love, support and encouragement in the past years. I am deeply grateful to my beloved wife Ngoc-Anh, who wholeheartedly took care of our small family, encouraged me, and together with me

overcome difficulties during staying in Japan. I thank my two sons, Anh-Chuong and Gia-Phuc, for bringing much energy and full of smiles to our small family. I would like to express special thanks to my parents for their endless love, sacrifice, and care to our brothers and sisters. I would like to thank my mother-in-law, who always encouraged me during my graduate studies. I also thank my brothers, Truong-Tho, Tuan-Anh, Truong-Thuat, and my sister Thu-Thuy, for always believing in me and encouraging me.

I would like to dedicate this dissertation to my parents,
to my wife Ngoc-Anh, and to my children Anh-Chuong and Gia-Phuc.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, the assessment in higher education has been shifting from traditional testing of asking only factual knowledge towards authentic assessment (Black and Wiliam, 1998; Dochy et al., 2006, 1999; Kvale, 2007). Authentic assessment aims at evaluating learner's proficiency in higher order skills and developed competencies (Jonsson and Svingby, 2007). In the context of authentic assessment, learning performance and learning activities are captured to evaluate such abilities by letting learners solve real-life, complex, and often open-ended assignments such as proving mathematical problems, developing program assignments, and writing reports (Jonsson and Svingby, 2007). However, when the number of learners increases as in Massive Open Online Courses (MOOCs), it is difficult for a few instructors to follow up every learner and individually assess assignments during the learning process (Capuano et al., 2017; Kulkarni et al., 2013; Sadler and Good, 2006). Instructor assessment is impossible to scale up to large classrooms or online courses with even thousands of simultaneous learners (Kulkarni et al., 2013; Piech et al., 2013).

One possible approach to overcome this assessment problem is to use computer-supported assessment tools (e.g., Paravati et al., 2017) to let the evaluation process can be done automatically (Capuano et al., 2017; Glance et al., 2013; Kulkarni et al., 2013). However, the variability of open-ended solutions of assignments and the lack of well-defined evaluation criteria interrupt reliable and valid assessment (Capuano et al., 2017; Kulkarni et al., 2013). Additionally, automated assessment cannot capture the semantics meaning of learning outcomes such as writing reports or design problems (Glance et al., 2013; Kulkarni et al., 2013; Paravati et al., 2017). This shortcoming limits the feedback that an automated assessment system can provide to help learners enhance learning (Kulkarni et al., 2013; Paravati et al., 2017).

A promising approach is peer assessment (Capuano et al., 2017; Kulkarni et al., 2013; Piech et al., 2013). Peer assessment, which is an assessment method based on a social constructivist approach, enables learners to assess outcomes or performance of their peers mutually (Dochy et al., 1999; Topping, 1998). Peer assessment provides many important learning benefits (Glance et al., 2013; Ueno and Okamoto, 2008; Uto and Ueno, 2016). It enables not only to give formative feedback to help learners enhance their learning (Dochy et al., 1999; Falchikov, 2005; Freeman, 1995; Lan et al., 2011; Lu and Law, 2012; Moccozet and Tardy, 2015; Papinczak et al., 2007; Staubitz et al., 2016; Topping, 1998) but also to provide summative assessments to estimate learner's ability (Capuano et al., 2017; Kulkarni et al., 2013; Piech et al., 2013). Moreover, when the number of learners increases, peer assessment can be conducted by dividing learners into multiple groups without burdening instructors and learners with assessment workload (Dochy et al., 1999; Kulkarni et al., 2013; Moccozet and Tardy, 2015; Piech et al., 2013; Sadler and Good, 2006; Sluijsmans et al., 2001; Suen, 2014). Therefore, peer assessment has been increasingly adopted in various large-scale e-learning and assessment situations (e.g., ArchMiller et al., 2016; Bhalerao and Ward, 2001; Davies, 2007; Lan et al., 2011; Lin et al., 2001; Sitthiworachart and Joy, 2004; Sung et al., 2005; Trahasch, 2004).

The accuracy of peer assessment, however, depends on rater characteristics such as rating severity and rating consistency (Sluijsmans et al., 2001; Ueno and Okamoto, 2008; Usami, 2010; Uto and Ueno, 2016; Wang and Yao, 2013). To solve this problem, several item response theory (IRT) models that incorporate rater characteristic parameters have been proposed (e.g., DeCarlo, 2005; Patz et al., 2002; Ueno et al., 2008; Usami, 2010; Uto and Ueno, 2016). Those IRT models provide more accurate ability assessment than the average/total scoring methods do because they can estimate the ability of learners considering rater characteristics (Uto and Ueno, 2016).

On the other hand, as mentioned above, when the number of learners increases as in MOOCs, peer assessment is often conducted by dividing learners into groups to alleviate the assessment workload of each learner. In this case, the accuracy of peer assessment also depends on the way to form groups (Nguyen et al., 2015; Wang and Yao, 2013).

To solve the problem, this study proposes a new group optimization method using IRT models with rater parameters and integer programming to maximize the accuracy of peer assessment conducted within each group. In particular, the proposed method is formulated as an integer programming problem to maximize the Fisher information, which is a widely used index to measure the accuracy of ability assessment in IRT.

However, experimental results reveal that, when peer assessment is conducted within each group, the proposed method cannot sufficiently improve the accuracy compared to the random group formation. The result suggests that it is difficult to assign raters with high Fisher information to all learners when peer assessment is conducted only within each group.

To address this limitation, this study introduces the concept of external raters for peer assessment conducted within each group and proposes an external rater selection method based on IRT models. In this study, an external rater is defined as a peer-rater who belongs to different groups. The proposed external rater selection method is formulated as an integer programming problem that maximizes the lower bound of the Fisher information of the estimated ability of the learners by the external raters. Experimental results using both simulated and real-world peer assessment data show that the introduction of external raters is useful to improve the accuracy sufficiently. Additionally, experimental results further demonstrate that the proposed external rater selection method sufficiently improves the accuracy of ability assessment in comparison to the random selection.

It is worth noting that several group formation methods have been proposed to support learners enhance their learning effectiveness in collaborative learning environments (e.g., Dascalu et al., 2014; Hübscher, 2010; Kardan and Sadeghi, 2016; Khandaker and Soh, 2010; Lin et al., 2016, 2010; Moreno et al., 2012; Ounnas et al., 2009; Pang et al., 2015; Sadeghi and Kardan, 2015; Srba and Bielikova, 2015). This study, however, does not examine the effectiveness of learning in collaborative learning environments. Nguyen et al. (2015) firstly attempted to address the problem of the accuracy of peer assessment conducted within groups. They proposed a method to form groups such that each learner is evaluated by as many peer-raters as possible to reduce the difference of accuracies of ability estimates among learners. However, that method does not guarantee the accuracy to be maximized.

Additionally, in the context of management area, several studies have also paid attention to the problem of using internal/external evaluations to assess the quality of training programs and organizations (e.g., Baartman et al., 2007; Bowen and Martens, 2006; Burke, 1998; Conley-Tyler, 2005; Lynn Snow et al., 2005; Nevo, 1994, 2001; Peavy et al., 2014; Ryan et al., 2007; Savoia et al., 2009; Shapiro et al., 2009; Torres et al., 1997; Volkov, 2011; Volkov and Baron, 2011; Withey et al., 1983; Wright et al., 2013). Those studies focus on the issues related to the reliability and objectivity of the internal/external evaluations and the impact of internal/external evaluations on improving organizational performance of those being evaluated. From the qualitative

analyses approach, the related literature suggests that external evaluations should be used for summative function of evaluation (Nevo, 1994), because of their reliability compared to internal evaluations (Conley-Tyler, 2005). External evaluators in those studies were defined as experts or professional evaluators who are not part of the target programs or organizations.

1.1 Outline of the Thesis

In Chapter 2, this study provides a review of group formation methods in the literature of collaborative learning. Recently, learning paradigm has remarkably shifted from individual learning towards collaborative learning. In more social and collaborative learning environments, learners can acquire more knowledge and transferable skills through learning together from the same situations. Collaborative learning is consistent with the constructivist approach proposed by Vygotsky (1978). Thus it has been broadly adopted in higher education as a pedagogic strategy to enhance individual learning. In the context of collaborative learning, forming learning groups is one of the challenging tasks. Chapter 2 therefore is devoted to review the recently advanced aspects related to the group formation problem in collaborative learning.

Chapter 3 provides a brief introduction to an e-learning management system called "Samurai" that this study uses to conduct peer assessment experiments. Next, this chapter defines rating data obtained from the peer assessment conducted within each group. Then this chapter introduces IRT with rater parameters for peer assessment. Finally, this chapter details the Fisher information, which is a widely adopted index to measure the accuracy of ability assessment in IRT.

Chapter 4 proposes a group optimization method using the IRT model and integer programming. The proposed group optimization method aims to maximize the accuracy of peer assessment conducted within each group. Concretely, the group optimization method is formulated as an integer programming problem that maximizes the lower bound of the Fisher information given to each learner. This chapter also examines several alternative objective functions to analyze the influence of objective functions related to the Fisher information on the performance of the proposed method. Next, this chapter presents experiments using simulated data to evaluate the performance of the proposed methods. Experimental results show that the groups formed by using the proposed methods cannot sufficiently improve the accuracy of ability assessment compared to the groups created randomly.

As an approach to overcome this limitation, Chapter 5 relaxes the constraint that restricts peer assessment to be conducted within each group only by introducing external raters. This chapter then proposes an external rater selection method to assign a few appropriate external raters to each learner after the proposed group optimization was conducted. This chapter formulates the external rater selection method as an integer programming problem that maximizes the lower bound of the Fisher information of the estimated ability of the learners given by the external raters. Then this chapter presents simulation experiments to demonstrate the effectiveness of the proposed method from three different perspectives. This chapter also describes experiments using real-world peer assessment data to demonstrate the effectiveness of the proposed method.

Finally, Chapter 6 summarizes the main contributions of this thesis, including (1) group optimization methods cannot sufficiently improve the accuracy of peer assessment conducted within each group compared to the random group formation, (2) the introducing of external raters to peer assessment is useful to enable improving the accuracy of ability assessment, and (3) the proposed external rater method can significantly improve the accuracy of peer assessment in comparison to the random rater selection.

Chapter 2

Related Work on Group Optimization

2.1 Introduction

Collaborative learning (CL) has been increasingly adopted in all levels of education (Strijbos, 2011) as a pedagogical strategy in which two or more learners in a group interact and learn together to accomplish a learning goal (Dillenbourg, 1999). Recently, with the introduction of computers into CL, Computer-Supported Collaborative Learning (CSCL) has emerged as a major field of research focusing on how technology can enhance CL (Chan and Van Aalst, 2004; Sadeghi and Kardan, 2015). CSCL environments provide learning situations where learners can participate in authentic activities (Chan and Van Aalst, 2004). Also, CSCL was designed based on social constructivist approaches (Vygotsky, 1978) to efficiently support students in representing, interpreting, and reflecting what they learned in knowledge-building communities (Chan and Van Aalst, 2004; Lin et al., 2016; Sadeghi and Kardan, 2015). Several studies indicate that CSCL provides a positive impact on promoting learner's motivation and on improving learning achievements (Lin et al., 2016; Sadeghi and Kardan, 2015).

In CL, one of the aspects that determines the productivity and the success of learning groups is the way to form groups (Sadeghi and Kardan, 2015; Seethamraju and Borman, 2009; Srba and Bielikova, 2015). Conventionally, the group formation process has employed random assignment, instructor-controlled grouping, or self-selected grouping methods (Hübscher, 2010; Lin et al., 2016; Srba and Bielikova, 2015). However, random assignment or self-selected grouping might create highly unbalanced groups (Lin et al., 2016; Srba and Bielikova, 2015). Instructor-controlled grouping can manage the unbalanced grouping problem. However, it is a relatively complicated

process and time-consuming, especially when the number of learners increases or an instructor does not understand students well (Srba and Bielikova, 2015). As a consequence, automatic group formation is one of the challenging problems and has attracted much interest of researchers (Hübscher, 2010; Lin et al., 2016, 2010; Moreno et al., 2012; Sadeghi and Kardan, 2015; Srba and Bielikova, 2015).

This chapter, therefore, is devoted to reviewing related work on the group formation methods in CL.

2.2 Group Formation in Collaborative Learning

2.2.1 Grouping algorithms

The most common approach to forming CL groups is to maximize diversity within groups (Hübscher, 2010). Diverse learning groups would provide positive effects on learning performance (e.g., Lin et al., 2016; Pang et al., 2015). For that purpose, Weitz and Lakshminarayanan (1998) formulated the maximum diversity student work-group problem, which now is known as the maximally diverse grouping problem (MDGP) (Brimberg et al., 2015). The MDGP creates groups to maximize the difference between pairwise students across all groups (Baker and Powell, 2002; Hübscher, 2010). The difference between two students can be defined by the summation of weighted contributions of grouping criteria from that the two students differ (Weitz and Lakshminarayanan, 1998) or by a distance function (e.g., Euclidean distance) between two students (Brimberg et al., 2015). Further detail of the calculation of the difference between students can be referred to Baker and Powell (2002), Gallego et al. (2013), Rodriguez et al. (2013), and Pang et al. (2015).

However, the MDGP is a NP-hard problem (Brimberg et al., 2015; Feo and Khellaf, 1990). Therefore, several heuristics algorithms to solve the problem have been proposed (e.g., Brimberg et al., 2015; Gallego et al., 2013; Rodriguez et al., 2013). Additionally, when applying the MDGP, what criteria should be considered to create productive CL groups is still an open research issue (Hübscher, 2010; Lin et al., 2016; Srba and Bielikova, 2015). Huxham and Land (2000) and Pang et al. (2015) have reported that there was no any evidence of the monotonic positive relationship between learning performance and diversities in demographics, personalities, and learning styles.

Because mathematical constraint models such as the MDGP are challenging to solve (Sadeghi and Kardan, 2015), other existing approaches resort to heuristic algorithms to form CL groups. A review of the literature reveals that evolutionary and swarm

intelligence algorithms have been widely adopted to form heterogeneous, homogeneous, and mixed groups (e.g., Dascalu et al., 2014; Gogoulou et al., 2007; Graf and Bekele, 2006; Lin et al., 2016, 2010; Moreno et al., 2012; Wang et al., 2007; Yannibelli and Amandi, 2011; Zheng and Pinkwart, 2014).

Clustering algorithms have also been used to solve the group formation problem. For example, fuzzy C-means clustering (Christodoulopoulos and Papanikolaou, 2007), K-means clustering (Ounnas et al., 2009; Pang et al., 2014), matrix-based clustering (Pollalis and Mavrommatis, 2009; Srba and Bielikova, 2015), hierarchical clustering (Zakrzewska, 2009), and hybrid clustering that combines fuzzy C-means and K-means algorithms (Montazer and Rezaei, 2012) have been proposed. Tanimoto (2007) employed the Squeaky Wheel algorithm to form groups that optimize the compatibility of a learner with the other peers in the same group. Herein, the compatibility denotes how much a learner would like to learn with peer-learners. Mahdi and Fattaneh (2013) proposed a modified Pareto Optimal Set (POS) algorithm called Semi-POS to form heterogeneous and homogeneous groups.

An agent-based approach has been employed to develop CL environments. Ikeda et al. (1997) and Inaba et al. (2000) developed a multi-agent system called FITS/CL. The system supports forming opportunistic groups so that the learning goal of each group member is consistent with the learning goal of the whole group. I-MINDS (Soh et al., 2008) learning system employs an iterative auction algorithm called VACAM (Soh et al., 2006) and a set of intelligent multi-agents to form groups with members who have high ability and social membership values. Recently, Khandaker and Soh (2010) also proposed a framework called iHUCOFS. That framework consists of multi-agents to help instructors form better groups over time by considering the evaluation of instructors as a grouping criterion in the next round of group formation.

Ounnas et al. (2009) have pointed out that the existing methods often fail in assigning all learners to groups, which was called as the orphan learner problem (Ounnas et al., 2009). As an approach to solving that problem, they first employed semantic web ontologies to model learner features dynamically. Then, they expressed the group formation problem as a constraint satisfaction problem given a set of constraints. Rubens et al. (2009) considered the group formation problem in informal CL environments without instructor's assistance, and learners are mainly self-directed. They proposed a method that automatically extracts information of learners from data sources such as academic publications or social networking sites and then forms CL groups.

More recently, Hübscher (2010) employed Tabu search algorithm to solve the constrained group formation problem related to general and context-specific criteria for project groups. Srba and Bielikova (2015) proposed an automatic formation of dynamic groups using group technology (GT) to create clusters of compatible learners based on the feedback obtained from the evaluation of previous collaborations. In that study, two learners are considered to be compatibility if their combination based on individual characteristics leads to positive learning achievement (Srba and Bielikova, 2015). Sadeghi and Kardan (2015) and Kardan and Sadeghi (2016) also formulated the group formation problem as a binary integer programming model to maximize the total "compatibility" between all individuals. That optimization model is as an extension of the *clique partitioning problem* (CPP) (Brimberg et al., 2017; Brusco and Köhn, 2009) applying to the group formation problem.

To enable forming groups with an arbitrary number of learner characteristics, Moreno et al. (2012) translated the group formation problem into a multi-objective optimization problem. They then employed genetic algorithms to demonstrate the effectiveness of the proposed method. Recently, Lin et al. (2016) have argued that the multi-objective grouping optimization problem related to learner characteristics should be considered as a trade-off between benefit objectives and cost objectives in CL, which often conflict with each other in optimization directions (Lin et al., 2016). To solve that problem, they proposed a trade-off multi-objective grouping optimization method that uses a technique for order preference by similarity to ideal solution (TOPSIS).

2.2.2 Grouping criteria

The review of the literature reveals that a variety of grouping criteria (i.e., learner characteristics) have been considered to form groups.

In general, grouping criteria include different aspects related to the learning status of learners. Learning knowledge was broadly adopted in several work to demonstrate the effectiveness of group formation methods (e.g., Brauer and Schmidt, 2012; Christodoulopoulos and Papanikolaou, 2007; Dascalu et al., 2014; Graf and Bekele, 2006; Lin et al., 2016; Moreno et al., 2012; Pang et al., 2015; Pollalis and Mavrommatis, 2009; Srba and Bielikova, 2015). Additionally, learning styles (e.g., Brauer and Schmidt, 2012; Christodoulopoulos and Papanikolaou, 2007; Huxham and Land, 2000; Montazer and Mohammad, 2013; Pang et al., 2015; Zakrzewska, 2009), level degree of interest or motivation (Dascalu et al., 2014; Graf and Bekele, 2006; Lin et al., 2010; Zakrzewska, 2009), skills and experiences (Brauer and Schmidt, 2012; Graf and Bekele, 2006; Hübscher, 2010), personal characteristics (Graf and Bekele, 2006; Ounnas et al.,

2009; Pang et al., 2015; Zheng and Pinkwart, 2014), thinking style (Wang et al., 2007), and context-specific preferences (Hübscher, 2010) were attempted and discussed.

Recently, social interactions (Brauer and Schmidt, 2012; Ounnas et al., 2009; Rubens et al., 2009) and the role of learners in a group (Ounnas et al., 2009; Yannibelli and Amandi, 2011) were also proposed.

2.3 Summary

This chapter has presented a literature review on the group formation methods to enhance CL in each group. The literature revealed that, in the context of CL, the group formation problem had been investigated mainly from two perspectives: (1) algorithms to help instructors create groups optimally under considered criteria, and (2) grouping criteria that effect to CL. Because of the increasing complexity of the problem both in many learners and criteria should be considered for the group formation problem, almost existing approaches resorted to heuristic algorithms to solve the problem.

The review of the grouping criteria highlighted a shortcoming that the existing methods have been facing a lack of standard metrics to enable measuring the quality of group formation processes. The existing methods have attempted to solve the problem from the context that the problem arose. Therefore, what characteristics should be considered to enable forming productive CL groups is still an open research issue.

Although it has acknowledged that assessment can strongly influence on CL (Lan et al., 2011; Sluijsmans and Strijbos, 2010; Strijbos, 2011), the current grouping optimization methods have paid much less attention to the perspective of assessment. In general, the assessment in CL is often focused on the final learning outcomes and is mainly conducted by instructors (Sluijsmans and Strijbos, 2010). Recently, Sluijsmans and Strijbos (2010) have argued that peer assessment is a suitable evaluation method for CL.

The literature review showed that before the present study, there was no study on group optimization methods to maximize the accuracy of peer assessment conducted within each group. In other words, applying existing group formation methods to optimize peer assessment groups does not guarantee the accuracy of ability assessment to be maximized.

Therefore, this study proposes a new group optimization method to maximize the accuracy of peer assessment.

Chapter 3

Item Response Theory for Peer Assessment

3.1 Introduction

Peer assessment, which enables learners to assess learning outcomes of their peers mutually (Dochy et al., 1999; Topping, 1998), has drawn much attention in recent years (ArchMiller et al., 2016; Capuano et al., 2017; Kulkarni et al., 2013; Lan et al., 2011; Strijbos, 2011; Suen, 2014; Uto and Ueno, 2016). Peer assessment provides many notable learning benefits (Glance et al., 2013; Ueno and Okamoto, 2008; Uto and Ueno, 2016), for instance:

1. Because assessment is integrated as a part of learning process, learning mistakes can be seen as learning opportunities rather than failures (Bostock, 2000).

2. Giving students rater role helps them improve learning motivation (Bostock, 2000; Weaver and Cotrell, 1986).

3. Learners can practice transferable skills such as evaluation and discussion skills (Bostock, 2000; Hamer et al., 2005).

4. Learners can learn from others' work and then induce self-reflection while they evaluate peers (Bostock, 2000; Hamer et al., 2005; Ueno and Okamoto, 2008).

5. Learners can receive readily understood feedback from other peers who have similar backgrounds (Ueno and Okamoto, 2008).

6. When the number of learners increases such as MOOCs, peer assessment can provide feedback to each learner without burdening instructor's workload (Shah et al., 2014; Suen, 2014).

7. As learners are mature adults, assessment results given by multiple raters are considered to be more reliable than those given by an instructor (Ueno and Okamoto, 2008).

Peer assessment, therefore, has been broadly adopted in many learning environments and evaluation situations (e.g., ArchMiller et al., 2016; Bhalerao and Ward, 2001; Cho and Schunn, 2007; Davies, 2007; Kulkarni et al., 2013; Lin et al., 2001; Sitthiworachart and Joy, 2004; Suen, 2014; Sung et al., 2005; Trahasch, 2004; Ueno and Okamoto, 2008; Uto and Ueno, 2016). In many e-learning environments, peer assessment has been mainly employed as a supportive learning tool to enrich individual learning by providing formative comments among learners (Lan et al., 2011; Lu and Law, 2012; Moccozet and Tardy, 2015; Papinczak et al., 2007). In recent years, peer assessment has also been increasingly adopted as a summative assessment tool to evaluate learner's ability such as in credential programs (Capuano et al., 2017; Kulkarni et al., 2013; Navrat and Tvarozek, 2014; Piech et al., 2013).

The accuracy of peer assessment, however, is known to depend on rater characteristics such as rating *severity* and rating *consistency* (Sluijsmans et al., 2001; Ueno and Okamoto, 2008; Usami, 2010; Uto and Ueno, 2016). As an approach to solving this problem, several item response theory (IRT) models incorporating rater characteristic parameters have been proposed. Previous studies have reported that those IRT models provide more accurate ability assessment than the average/total scoring methods do because they can estimate the ability of learners considering rater characteristics (Ueno et al., 2008; Usami, 2010; Uto and Ueno, 2016).

This chapter introduces an IRT model with rater characteristic parameters that this study employs. Firstly, this chapter briefs an introduction to a learning management system (LMS) called "*Samurai*", which is used as the peer assessment platform in this study. Then, this chapter formulates peer assessment data conducted within groups using the Samurai system. Next, this chapter explains the IRT model for peer assessment proposed by Uto and Ueno (2016). Finally, the detail of Fisher information, which is a widely adopted index to evaluate the accuracy of the ability assessment in IRT, is presented.

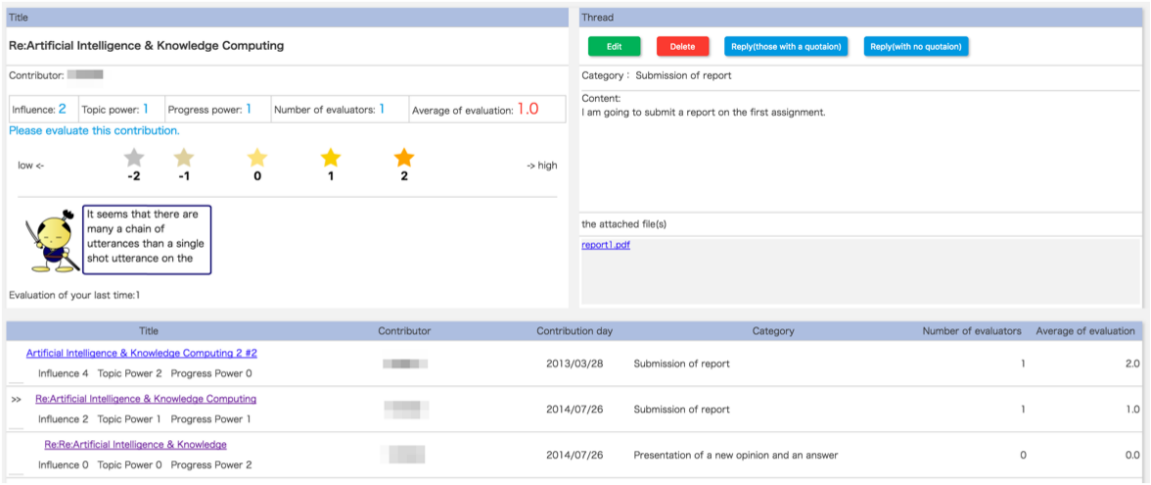


Figure 3.1 Peer assessment interface of the LMS Samurai.

3.2 Peer Assessment

3.2.1 Peer assessment platform

The LMS Samurai (Ueno, 2004) stores a large number of e-learning courses. Each course consists of 15 content sessions tailored for 90-min classes (with units are designated as topics). Each topic comprises instructional text screens, images, videos, and practice tests. How learners respond to the sessions and how long it takes them to complete the lesson are stored automatically in learning history database of the system. Those data are analyzed using various data mining techniques. The analysis results are used for facilitating learning.

In some courses, writing reports are assigned to learners. The Samurai system has a discussion board system that enables learners to submit reports and to conduct peer assessment among them. Figure 3.1 depicts an interface where a learner submits a report. The lower half of Figure 3.1 presents hyper-links to comments given by peer-learners. By clicking a hyper-link, detail of comments are displayed in the upper right of Figure 3.1. The top left shows five-star buttons used for assigning ratings. These buttons include -2 (Bad), -1 (Poor), 0 (Fair), 1 (Good), and 2 (Excellent). The learner who submitted the report can consider these ratings and comments to revise his/her work accordingly. The average rating score of the report is calculated from the peer assessment data and then is stored in the system. This score is often used to recommend excellent reports to the other learners in the system (Ueno and Uto, 2011). This score has also been used in various purposes, such as grading learners (e.g., Capuano et al., 2017; Dochy et al., 1999; Sadler and Good, 2006), evaluating

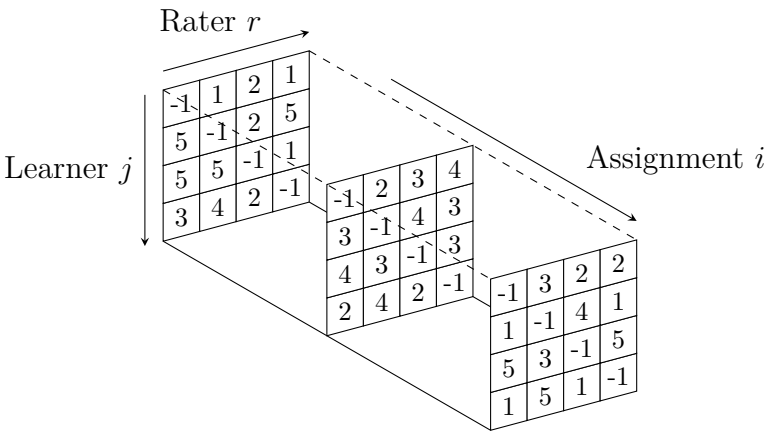


Figure 3.2 An example of peer assessment data.

rater reliability (e.g., Piech et al., 2013), and assigning weights to formative comments (e.g., Suen, 2014). This study aims to improve the accuracy of this rating score.

3.2.2 Peer assessment data

The rating data $\boldsymbol{U}$ obtained from the described peer assessment system above consist of rating categories $k \in \boldsymbol{K} = \{1, \ldots, K\}$ given to each learning outcome of learner $j \in \boldsymbol{J} = \{1, \ldots, J\}$ by each peer-rater $r \in \boldsymbol{J}$ for each assignment $i \in \boldsymbol{N} = \{1, \ldots, N\}$. Let u_{ijr} be a response of rater r to learner j's outcome for assignment i, the data $\boldsymbol{U}$ are formulated as follows.

$$\boldsymbol{U} = \{u_{ijr} \mid u_{ijr} \in \boldsymbol{K} \cup \{\text{-1}\}, i \in \boldsymbol{N}, j \in \boldsymbol{J}, r \in \boldsymbol{J}\}, \tag{3.1}$$

which $u_{ijr} = -1$ denotes missing data. This study uses five categories $\{1, 2, 3, 4, 5\}$ transformed from the rating buttons $\{-2, -1, 0, 1, 2\}$ in the system above. Figure 3.2 depicts an example of peer assessment data. These data are three-way data since they comprise of learners $\times$ raters $\times$ assignments.

As introduced in Chapter 1, when the number of learners increases, peer assessment is often conducted by dividing learners into multiple groups to reduce learners' assessment workload. This study assumes that learning groups are formed for each assignment $i \in \boldsymbol{N}$. Thus, let

$$x_{igjr} = \begin{cases} 1, & \text{if learner } j \text{ and peer-rater } r \text{ are in the same group } g \text{ on assignment } i, \\ 0, & \text{otherwise.} \end{cases}$$

Then, the groups of peer assessment for assignment i can be formulated as follows.

$$\boldsymbol{X}_i = \{x_{igjr} \mid x_{igjr} \in \{0,1\}, i \in \boldsymbol{N}, g \in \boldsymbol{G}, j \in \boldsymbol{J}, r \in \boldsymbol{J}\}. \tag{3.2}$$

When peer assessment is conducted within each group only, the rating data u_{ijr} become missing data if two learner j and r do not belong to the same group (i.e., $\sum_{g \in \boldsymbol{G}} x_{igjr} = 0$).

This study aims to improve the accuracy of ability assessment obtained from the peer assessment data $\boldsymbol{U}$ by optimizing the group formation $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$. For that purpose, this study uses item response theory.

3.3 Item Response Theory

Item response theory (IRT) (Lord, 1980), which is a test theory based on mathematical models, has been widely adopted in many areas of educational testing. IRT models define the probability that a learner responds to a test item as a function of the latent ability of the learner and item characteristics (e.g., difficulty and discrimination). IRT models offer many benefits, for instance (Ueno and Okamoto, 2008; Uto and Ueno, 2016):

1. It is possible to estimate learner ability while minimizing the effects of different or aberrant items that lead to low measurement accuracy.

2. The learner's responses to various test items can be evaluated on the same scale.

3. It is easy to handle missing data.

Conventionally, IRT models such as Rash model (Rasch, 1966), two-parameter logistic (2PL) model (Lord, 1980) have been applied to test items for which the responses can be scored automatically as correct or wrong, such as multiple-choice items. In recent years, several polytomous IRT models have also proposed to apply to performance assessment such as essay written tests (DeCarlo, 2005; Matteucci and Stracqualursi, 2006; Muraki et al., 2000).

Several well-known polytomous IRT models include Rating Scale Model (RSM) (Andrich, 1978), Partial Credit Model (PCM) (Masters, 1982), Generalized Partial Credit Model (GPCM) (Muraki, 1992) and Graded Response Model (GRM) (Samejima, 1969). The following subsection introduces the GRM, which is the fundamental model of an IRT model extended for peer assessment that this study uses.

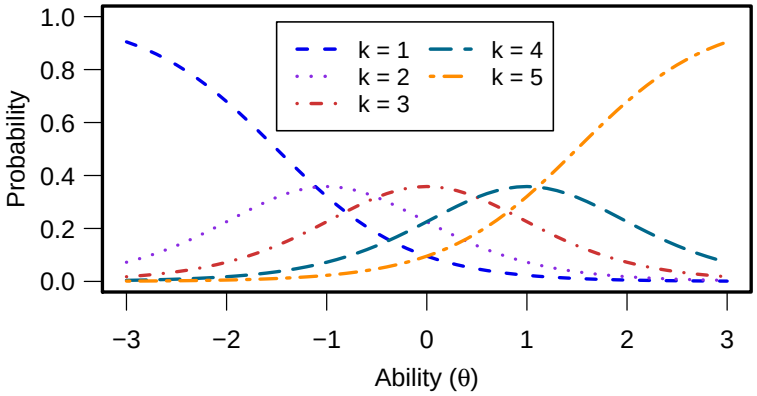


Figure 3.3 Item characteristic curves of the graded response model for five categories.

3.3.1 Grade Response Model

The GRM model defines the probability that learner j responds to category k of item i as follows.

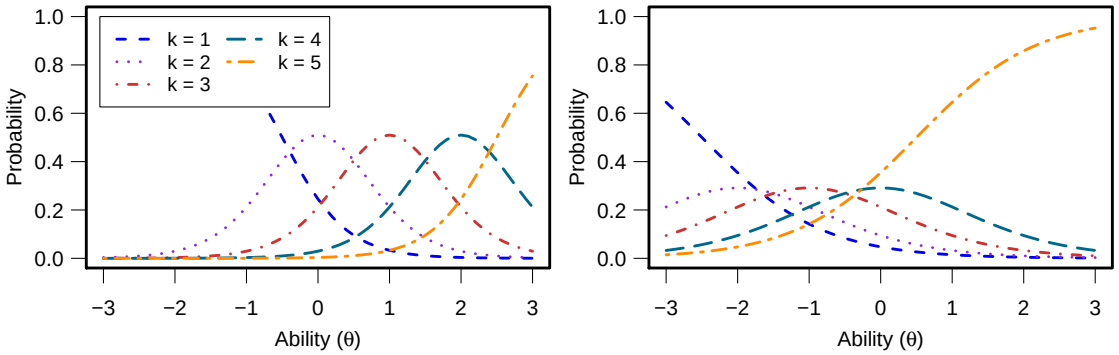
$$P_{ijk} = P^*_{ij,k-1} - P^*_{ijk}, \tag{3.3}$$

$$\begin{cases} P^*_{ij0} = 1, \\ P^*_{ijk} = [1 + \exp(-\alpha_i(\theta_j - \beta_{ik}))]^{-1}, \ k = 1, \ldots, K-1, \\ P^*_{ijK} = 0. \end{cases} \tag{3.4}$$

Here, parameter α_i indicates the discrimination of item i, parameter β_{ik} represents the difficulty in obtaining the score k of item i, and parameter θ_j denotes the ability level of learner j. In this model, the order of the difficulty parameters is restricted to $\beta_{i1} < \cdots < \beta_{i,K-1}$.

Figure 3.3 depicts an example of item response curves of the GRM model with $K = 5$, $\alpha_i = 1.5$, $\beta_{i1} = -1.5$, $\beta_{i2} = -0.5$, $\beta_{i3} = 0.5$, and $\beta_{i4} = 1.5$. The horizontal axis denotes the ability level θ, and the vertical axis presents the probability that a learner with ability level θ responses to category k. Figure 3.3 shows that learners with lower (higher) ability level tend to respond in lower (higher) categories.

Traditional IRT models such as GRM are assumed to be applied to two-way data that consists of learners $\times$ items. However, as described in Section 3.2.2, peer assessment data $\boldsymbol{U}$ are three-way data consisting of learners $\times$ raters $\times$ assignments. Consequently, traditional IRT models are not capable of applying to these three-way data directly.



(a) Rater 1: $\alpha_r = 1.5$, $\epsilon_r = 1.0$ (b) Rater 2: $\alpha_r = 0.8$, $\epsilon_r = -1.0$

Figure 3.4 Item characteristic curves for two different raters for five categories.

Recently, as an approach to solving that problem, several studies have proposed IRT models that incorporate rater characteristic parameters (DeCarlo, 2005; Patz and Junker, 1999; Ueno and Okamoto, 2008; Usami, 2010; Uto and Ueno, 2016). In those models, characteristic parameters of items are considered as characteristic parameters of assignments. Those models can accurately estimate learner ability level considering rater characteristics. The next subsection introduces an IRT model proposed by Uto and Ueno (2016) for peer assessment, which is known to provide the highest accuracy of ability assessment in the relevant models when the number of peer-raters increases.

3.3.2 Item Response Theory for Peer Assessment

Uto and Ueno (2016) have proposed a GRM that incorporates rater characteristic parameters for peer assessment. The model defines the probability that rater r responds to learner j's outcome in the category k of assignment i as follows.

$$P_{ijrk} = P(u_{ijr} = k \mid \theta_j) = P^*_{ijr,k-1} - P^*_{ijrk} \tag{3.5}$$

$$\begin{cases} P^*_{ijr0} = 1, \\ P^*_{ijrk} = [1 + \exp(-\alpha_i \alpha_r (\theta_j - \beta_{ik} - \varepsilon_r))]^{-1}, \ k = 1, \ldots, K-1, \\ P^*_{ijrK} = 0. \end{cases} \tag{3.6}$$

In this model, parameters α_r and ε_r reflect the consistency and severity of rater r; parameter α_i indicates the discrimination of assignment i; and parameter β_{ik} presents the difficulty in obtaining category k for assignment i (with constraint $\beta_{i1} < \cdots < \beta_{i,K-1}$). Additionally, $\alpha_{r=1} = 1$ and $\varepsilon_{r=1} = 0$ are assumed to identify the model.

To explain the effects of rater parameters, Figure 3.4 shows item characteristic curves of two raters with assignment parameters $\alpha_i = 1.5$, $\beta_{i1} = -1.5$, $\beta_{i2} = -0.5$, $\beta_{i3} = 0.5$, and $\beta_{i4} = 1.5$. In this example, the number of categories K was set to five. The left panel presents item characteristic curves of *Rater* 1, who has $\alpha_r = 1.5$ and $\epsilon_r = 1.0$. The right panel shows item characteristic curves of *Rater* 2, who has $\alpha_r = 0.8$ and $\epsilon_r = -1.0$. In Figure 3.4, the horizontal axis denotes learner ability level θ, and the vertical axis shows the probability of rating responses to each category.

According to Figure 3.4, the higher the rater consistency parameter is, the larger the differences in the response probability among the rating categories are. It means that a rater whose a higher consistency can distinguish the differences in performance of each learner more accurately and consistently. Additionally, Figure 3.4 shows that the item response function of *Rater* 1, who has higher severity, shifted to the right compared to those of *Rater* 2. Namely, a higher performance is necessary to obtain a score from *Rater* 1 than to obtain the same score from *Rater* 2.

The IRT models with rater parameters such as the model presented above are possible to estimate learner's ability more accurately than the average scoring method because they can estimate the learner abilities considering the influence of rater characteristics (Uto and Ueno, 2016). Furthermore, the ability values obtained by applying IRT models incorporating rater characteristic parameters to peer assessment data is known more accurately than the results obtained from the assessment data given by an instructor only (Ueno and Okamoto, 2008). Recently, the ability values obtained from peer assessment has been increasingly used for various purposes, for instance, learner's grading judgment (Capuano et al., 2017; Kulkarni et al., 2013; Sadler and Good, 2006; Sluijsmans et al., 2001), ability judgment (Piech et al., 2013), and recommending excellent learning outcomes of other learners (Ueno, 2004). Therefore, improving the accuracy of peer assessment is essential.

The unique feature of the IRT model proposed by Uto and Ueno (2016) is that each rater has only one consistency and severity parameter respectively. As a result, when the number of raters increases, the number of rater parameters in the model increases more slowly than those in conventional models that incorporate higher dimensional rater parameters (Uto and Ueno, 2016). The accuracy of parameter estimation is known to be higher if a model has fewer parameters when the number of data per parameter increases (Bishop, 2006; Uto and Ueno, 2016). This study assumes that peer assessment conducting within each group is necessary because of the increasing number of learners (= raters). In this case, the Uto and Ueno (2016) model can provide better

performance than the similar models proposed previously does. Therefore, the present study adopts this model.

3.3.3 Fisher information

Let $\hat{\theta}$ be the estimated value of the ability parameter for a learner with truth ability level θ. The variance of $\hat{\theta}$ given θ, which is denoted as $\mathrm{Var}(\hat{\theta} \mid \theta)$, over replications of the assessment is considered as an appropriate measurement for the accuracy of the ability estimation (Van der Linden, 2006).

In IRT, the variance function of any unbiased estimator $\hat{\theta}$ is asymptotically equal to the inverse of the Fisher information, which is often denoted as $I(\theta)$ (Lord, 1980). According to the *Cramér-Rao inequality* (Frieden, 2004; Lord, 1980), this relation can be written as

$$\mathrm{Var}(\hat{\theta} \mid \theta) \geq \frac{1}{I(\theta)}. \tag{3.7}$$

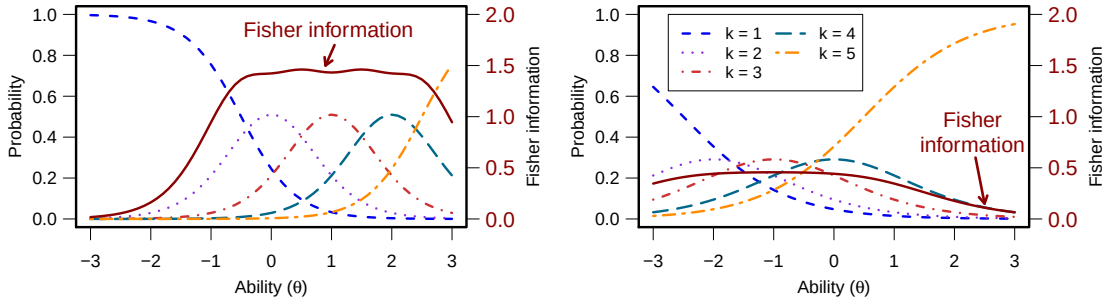
A higher value of the Fisher information implies smaller variance of ability estimates. Namely, a higher value of the Fisher information provides better accuracy of ability assessment. Thus, the Fisher information has been widely used as an index to measure the accuracy of the ability estimates.

For the model proposed by Uto and Ueno (2016), the Fisher information when rater r assesses an outcome of learner j with ability level θ_j on assignment i can be calculated as follows.

$$\begin{aligned} I_{ir}(\theta_j) &= -\,\mathbb{E}\left[\frac{\partial^2}{\partial \theta_j^2} \log P_{ijrk}\right] \\ &= \alpha_i^2 \alpha_r^2 \sum_{k=1}^{K} \frac{\left(P^*_{ijr,k-1} Q^*_{ijr,k-1} - P^*_{ijrk} Q^*_{ijrk}\right)^2}{P^*_{ijr,k-1} - P^*_{ijrk}}, \end{aligned} \tag{3.8}$$

with $Q^*_{ijrk} = 1 - P^*_{ijrk}$.

Figure 3.5 depicts an example of the Fisher information given by the two different raters that have been explained in Subsection 3.3.2 using Uto and Ueno (2016) model with assignment parameters $\alpha_i = 1.5$, $\beta_{i1} = -1.5$, $\beta_{i2} = -0.5$, $\beta_{i3} = 0.5$, and $\beta_{i4} = 1.5$. In this example, the number of categories $K = 5$ was used. The left panel presents the Fisher information given by *Rater* 1, who has $\alpha_r = 1.5$ and $\epsilon_r = 1.0$. The right panel shows the Fisher information given by *Rater* 2, who has $\alpha_r = 0.8$ and $\epsilon_r = -1.0$. In Figure 3.5, the horizontal axis denotes learner ability level θ. The left vertical axis shows the probability of rating responses to each category and the right vertical axis presents the Fisher information values corresponding to that response probability.



(a) Rater 1: $\alpha_r = 1.5$, $\epsilon_r = 1.0$ (b) Rater 2: $\alpha_r = 0.8$, $\epsilon_r = -1.0$

Figure 3.5 An example of the Fisher information given by two different raters.

According to Figure 3.5, the Fisher information given by *Rater* 1, who can accurately evaluate the performance of each learner, is higher than the corresponding values given by *Rater* 2. Furthermore, *Rater* 1, who is more severe than *Rater* 2, provides higher Fisher information to learners with ability above the average compared to *Rater* 2 in the same ability range. *Rater* 2, who is extremely lenient rater, however, gives higher Fisher information to learners with the ability bellow the average in comparison with *Rater* 1.

An attractive property of the Fisher information functions is that they are additive (Lord, 1980; Van der Linden, 2006). Thus, when peer assessment is conducted within each group, the information for learner j on assignment i can be defined by the summation of the information given by each peer-rater in the same group.

$$I_i(\theta_j) = \sum_{\substack{r \in \boldsymbol{J} \\ r \neq j}} \sum_{g \in \boldsymbol{G}} I_{ir}(\theta_j) x_{igjr}. \tag{3.9}$$

This study does not consider self-assessment. Therefore, in equation (3.9) above, the constraint $r \neq j$ is given.

For all of assignments $i \in \boldsymbol{N} = \{1, \ldots, N\}$, the Fisher information function becomes

$$I(\theta_j) = \sum_{i \in \boldsymbol{N}} I_i(\theta_j). \tag{3.10}$$

A higher Fisher information means that the assigned peer-raters would more accurately assess the ability level θ_j of learner j.

3.4 Summary

This chapter presented the peer assessment platform used in this study and an IRT model for peer assessment. The Fisher information, which is an index of ability assessment accuracy, was also explained in detail.

The accuracy of peer assessment is expected to be improved if the IRT models incorporating rater characteristic parameters are employed to estimate the ability parameters. However, when peer assessment is conducted within each group, the accuracy of ability assessment also depends on how to form groups (Nguyen et al., 2015; Wang and Yao, 2013). In this case, a group optimization considering rater characteristics is required to improve the accuracy of ability assessment.

The following chapter proposes a new group optimization method.

Chapter 4

Group Optimization using Item Response Theory

4.1 Introduction

As stated in the previous chapter, an optimization of groups considering rater characteristics is required to improve the accuracy of ability assessment when peer assessment is conducted within groups. However, the literature review revealed that only Nguyen et al. (2015) firstly drawn an attempt to address the problem. In that study, they proposed a method to form groups so that each learner is assessed by as many peer-raters as possible to reduce the difference of accuracies of ability estimates among learners. However, that method does not maximize the accuracy of peer assessment.

To solve the problem, this chapter proposes a new group optimization method to maximize the accuracy of ability assessment using IRT models for peer assessment. As presented in Subsection 3.3.3, the accuracy of peer assessment would be maximized if the Fisher information given by peer-raters to each learner in each group is maximized. Therefore, this study proposes a group optimization method that maximizes the Fisher information given to each learner.

4.2 Group Optimization based on IRT

This section formulates the group optimization problem using IRT models that incorporate rater characteristic parameters as an integer programming problem. In this study, the groups are optimized for each assignment $i \in \boldsymbol{N} = \{1, \ldots, N\}$.

The group optimization method for assignment i based on IRT models that incorporate rater characteristic parameters is formulated as the following integer programming problem.

$$\begin{aligned}
&\text{maximize} && y_i && && (4.1)\\
&\text{subject to} && && && \\
& && \sum_{\substack{r\in\boldsymbol{J}\\ r\neq j}}\sum_{g\in\boldsymbol{G}} I_{ir}(\theta_j)x_{igjr} \geq y_i, && \forall j, && (4.2)\\
& && \sum_{g\in\boldsymbol{G}} x_{igjj} = 1, && \forall j, && (4.3)\\
& && \sum_{g\in\boldsymbol{G}}(1-x_{igjj})\sum_{r\in\boldsymbol{J}} x_{igjr} = 0, && \forall j, && (4.4)\\
& && n_l \leq \sum_{j\in\boldsymbol{J}} x_{igjj} \leq n_u, && \forall g, && (4.5)\\
& && n_l \leq \sum_{g\in\boldsymbol{G}} x_{igjj}\sum_{r\in\boldsymbol{J}} x_{igjr} \leq n_u, && \forall j, && (4.6)\\
& && x_{igjr} = x_{igrj}, && \forall g,j,r, && (4.7)\\
& && x_{igjr} \in \{0,1\}, && \forall g,j,r. && (4.8)
\end{aligned}$$

In the formulated problem above, constraints (4.2) restrict that the Fisher information given to each learner j must be greater than or equal to the lower bound y_i. Constraints (4.3) and (4.4) ensure that each learner is assigned to only one group for each assignment i. The constraints in (4.5) and (4.6) control the number of learners assigning to each group. Herein, parameters n_l and n_u respectively denote the lower bound and upper bound of the number of learners in a group. This study uses conditions $n_l = \lfloor J/G \rfloor$ and $n_u = \lceil J/G \rceil$ to equalize the number of learners among groups, which the symbols $\lfloor\ \rfloor$ and $\lceil\ \rceil$ respectively denote floor and ceiling functions. Constraints (4.7) assure that if learner j and learner r are in the same group g, they must assess each other.

The objective function in (4.1) aims at maximizing the value y_i for each assignment i. In other words, the proposed group optimization problem maximizes the lower bound of the Fisher information given to each learner. This optimization model, therefore, is also called *maximin* optimization (Adema, 1989).

By solving the problem, we can obtain groups that the Fisher information given to each learner was maximized as much as possible.

4.2.1 Alternative objective functions

The objective function of the formulated optimization problem maximizes the lower bound of the Fisher information given to each learner. However, other objective functions can also be employed to maximize the Fisher information given to each learner. This subsection considers a variety of plausible alternatives.

To distinguish from other alternatives, the objective function in the formulated optimization problem is called as the Z_1 function.

$$\begin{aligned} &\text{maximize} && y_i \\ &\text{subject to} \\ & && Z_1 \coloneqq \sum_{\substack{r \in \boldsymbol{J} \\ r \neq j}} \sum_{g \in \boldsymbol{G}} I_{ir}(\theta_j) x_{igjr} \geq y_i, && \forall j. \end{aligned}$$

The first alternative defines an objective function that maximizes the total amount of the Fisher information given to each learner. Thus, the objective function would be formulated as follows.

$$\begin{aligned} &\text{maximize} && y_i \\ &\text{subject to} \\ & && Z_2 \coloneqq \sum_{j \in \boldsymbol{J}} \sum_{\substack{r \in \boldsymbol{J} \\ r \neq j}} \sum_{g \in \boldsymbol{G}} I_{ir}(\theta_j) x_{igjr} = y_i. && (4.9) \end{aligned}$$

The second possible alternative objective function is to maximize the lower bound of the Fisher information given to each group. Concretely, the objective function can be defined as the following equation.

$$\begin{aligned} &\text{maximize} && y_i \\ &\text{subject to} \\ & && Z_3 \coloneqq \sum_{j \in \boldsymbol{J}} \sum_{\substack{r \in \boldsymbol{J} \\ r \neq j}} I_{ir}(\theta_j) x_{igjr} \geq y_i, && \forall g. && (4.10) \end{aligned}$$

4.3 Evaluation using simulated data

In the proposed group optimization method, learners who can accurately evaluate each other are assigned to the same group. The method, therefore, is expected to improve the accuracy of ability assessment.

Table 4.1 Prior distributions for the IRT model with rater parameters.

$\theta_j \sim N(0.0, 1.0)$
$\log \alpha_r \sim N(0.0, 0.5)$, $\epsilon_r \sim N(0.0, 0.8)$
$\log \alpha_i \sim N(0.1, 0.4)$, $\beta_{ik} \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
$\boldsymbol{\mu} = (-2.0, -0.75, 0.75, 2.0)$
$\boldsymbol{\Sigma} = \begin{pmatrix} 0.16 & 0.10 & 0.04 & 0.04 \\ 0.10 & 0.16 & 0.10 & 0.04 \\ 0.04 & 0.10 & 0.16 & 0.10 \\ 0.04 & 0.04 & 0.10 & 0.16 \end{pmatrix}$

This section evaluates the performance of the proposed method. Concretely, this study conducted the following simulation experiment.

1. For $J \in \{15, 30\}$ and $N \in \{4, 5\}$, the true parameters of the IRT model described in Section 3.3.2 were generated randomly from the prior distributions in Table 4.1. The values of J and N were employed to meet the situations of two actual e-learning courses data collected from the Samurai system from 2007 to 2013. More specifically, the condition $J \in \{15, 30\}$ was employed because the average number of learners in each course was 12.9 (standard deviation = 4.2) and 32.9 (standard deviation = 14.6), respectively. And the condition $N \in \{4, 5\}$ was used because the number of assignments in each course was four and five.

2. For each assignment i, learners were divided into G groups using the proposed method (designated as *MxFiG* with objective functions Z_1–Z_3) and a random group formation method (designated as *RndG*). The number of groups is usually determined so that each group has from 3 to 14 members (Cho et al., 2016; Lin et al., 2016; Papinczak et al., 2007; Sluijsmans et al., 2001). In this study, $G \in \{3, 4, 5\}$ for $J = 15$ and $G \in \{3, 4, 5, 10\}$ for $J = 30$ were set because the number of group members falls within this range when $J \in \{15, 30\}$. The proposed method was solved using *IBM ILOG CPLEX Optimization Studio* (IBM Corp., 2015). A feasible solution is employed if the optimal solution could not be found within five minutes. Additionally, for the proposed method, the Fisher information was calculated using the true parameters to evaluate the performance in the ideal conditions.

3. Given the constructed groups and the true parameters, rating data were sampled randomly based on the IRT model.

4. The ability of learners was estimated from the sampled rating data given the true parameters of raters and assignments. The expected a posteriori (EAP) estimation method using Gaussian quadrature was employed to estimate (Baker and Kim, 2004).

5. The root mean square deviation (RMSE) between the estimated ability and the true ability was calculated using the following equation:

$$\mathrm{RMSE} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(\hat{\theta}_j - \theta_j)^2}. \tag{4.11}$$

   Here, $\hat{\theta}_j$ and θ_j are the estimated ability and the true ability of learner j respectively. The Fisher information given to each learner and each group was also calculated.

6. After repeating the procedures 1–5 above 10 times, the mean and standard deviation of the RMSE and Fisher information values were calculated.

The mean values of the Fisher information given to each learner and RMSE are presented in Table 4.2 and Table 4.3 , respectively. The values of standard deviation of the Fisher information given to each group are shown in Table 4.4.

The results show that the Fisher information increases and the RMSE values decrease when the number of assignments N increases or the number of groups G decreases because, in that cases, the number of rating data given to each learner increases. This is a direct consequence of the result explained in inequality (3.7), and equations (3.9), (3.10). This result is also consistent with the results reported in (Uto and Ueno, 2016). Uto and Ueno (2016) showed that in general, the increasing of rating data for each learner improves the ability assessment accuracy.

According to Table 4.2, the proposed method with three objective functions Z_1–Z_3 provided higher Fisher information than the random grouping method did in all cases.

However, the RMSE values in Table 4.3 show that the proposed method could not sufficiently improve the accuracy of ability assessment compared to the random method. It can be explained that because the improvement of the Fisher information given by the proposed method was small and that improvement was not enough to sufficiently improve the accuracy.

Comparing among objective functions, the objective function Z_1 provided better performance than the other ones. The objective function Z_2 considerably improved the average value of the Fisher information compared to the Z_1 and Z_3 functions. However,

Table 4.2 Fisher information of grouping methods using simulated data.

(a) $J = 15$

			MxFiG		
N	G	RndG	Z_1	Z_2	Z_3
4	3	9.182	**9.604**	**10.285**	**9.814**
		(2.370)	(2.671)	(2.978)	(2.695)
	4	6.355	**6.426**	**7.670**	**6.662**
		(1.710)	(1.814)	(2.290)	(1.866)
	5	4.604	**4.780**	**5.334**	**4.853**
		(1.202)	(1.308)	(1.605)	(1.335)
-	-	-	-	-	-
-	-	-	-	-	-
5	3	11.156	**11.671**	**12.455**	**11.891**
		(2.570)	(2.984)	(3.182)	(2.924)
	4	7.781	**7.826**	**9.281**	**8.092**
		(1.766)	(2.040)	(2.443)	(2.100)
	5	5.454	**5.801**	**6.450**	**5.908**
		(1.216)	(1.421)	(1.714)	(1.492)
-	-	-	-	-	-
-	-	-	-	-	-

(b) $J = 30$

			MxFiG		
N	G	RndG	Z_1	Z_2	Z_3
4	3	15.919	**16.227**	**17.560**	**17.123**
		(4.592)	(4.741)	(5.982)	(5.195)
	4	11.546	**11.844**	**13.256**	**12.421**
		(3.277)	(3.524)	(4.324)	(3.848)
	5	8.767	**9.169**	**10.056**	**9.533**
		(2.547)	(2.774)	(3.322)	(2.867)
	10	3.501	**3.599**	**4.130**	**3.725**
		(1.019)	(1.029)	(1.401)	(1.105)
5	3	20.340	**20.872**	**22.489**	**21.965**
		(5.110)	(5.345)	(6.546)	(5.778)
	4	14.822	**15.195**	**16.971**	**15.951**
		(3.756)	(3.934)	(4.727)	(4.260)
	5	11.356	**11.718**	**12.881**	**12.251**
		(2.884)	(3.066)	(3.624)	(3.193)
	10	4.518	**4.644**	**5.292**	**4.786**
		(1.115)	(1.186)	(1.522)	(1.247)

Table 4.3 RMSE of grouping methods using simulated data.

(a) $J = 15$

			MxFiG		
N	G	RndG	Z_1	Z_2	Z_3
4	3	**0.315**	0.337	0.344	0.325
		(0.084)	(0.054)	(0.088)	(0.071)
	4	0.399	**0.396**	0.404	0.408
		(0.091)	(0.094)	(0.088)	(0.120)
	5	0.466	0.447	**0.437**	0.451
		(0.109)	(0.090)	(0.150)	(0.090)
-	-	-	-	-	-
-	-	-	-	-	-
5	3	0.310	0.313	0.298	**0.287**
		(0.080)	(0.084)	(0.081)	(0.076)
	4	**0.333**	0.356	0.359	0.369
		(0.078)	(0.099)	(0.080)	(0.114)
	5	0.395	0.413	**0.378**	0.464
		(0.100)	(0.094)	(0.105)	(0.113)
-	-	-	-	-	-
-	-	-	-	-	-

(b) $J = 30$

			MxFiG		
N	G	RndG	Z_1	Z_2	Z_3
4	3	0.261	**0.227**	0.257	0.250
		(0.039)	(0.046)	(0.055)	(0.060)
	4	**0.268**	0.292	0.297	0.311
		(0.038)	(0.048)	(0.049)	(0.044)
	5	**0.310**	0.336	0.318	0.326
		(0.051)	(0.068)	(0.042)	(0.059)
	10	0.494	**0.466**	0.484	0.539
		(0.042)	(0.077)	(0.096)	(0.069)
5	3	0.218	**0.212**	0.219	0.216
		(0.033)	(0.042)	(0.048)	(0.040)
	4	**0.246**	0.254	0.258	0.266
		(0.042)	(0.037)	(0.054)	(0.038)
	5	0.299	0.288	**0.282**	0.298
		(0.056)	(0.052)	(0.041)	(0.039)
	10	0.431	**0.409**	0.432	0.458
		(0.057)	(0.072)	(0.089)	(0.073)

Table 4.4 Fisher information of each group using simulated data.

(a) $J = 15$

			MxFiG		
N	G	RndG	Z_1	Z_2	Z_3
4	3	47.400	53.438	59.569	53.912
	4	25.655	27.221	34.352	27.998
	5	14.434	15.706	19.266	16.025
-	-	-	-	-	-
5	3	64.269	74.604	79.571	73.112
	4	33.122	38.264	45.808	39.383
	5	18.245	21.322	25.712	22.381
-	-	-	-	-	-

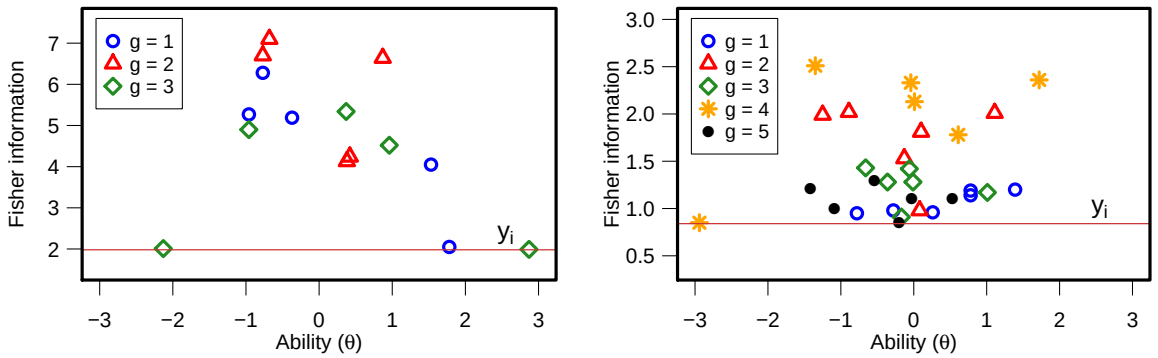
(b) $J = 30$

			MxFiG		
N	G	RndG	Z_1	Z_2	Z_3
4	3	183.712	189.665	221.144	207.815
	4	98.327	105.730	129.744	115.453
	5	61.142	66.585	79.750	68.813
	10	12.238	12.356	16.814	13.268
5	3	255.527	267.285	304.745	288.928
	4	140.863	147.545	177.267	159.764
	5	86.523	91.989	108.735	95.790
	10	16.735	17.792	22.830	18.705

the objective function Z_2 tends to form unbalanced groups, which some learners are given an extremely high Fisher information and others are given a small Fisher information. Because maximizing the summation of the Fisher information given to each learner leads to retaining peer-raters who provide the Fisher information with large values and cutting the ones who give small values as much as possible. The values of standard deviation of the Fisher information given to each leaner shown in Table 4.2 demonstrate this argument. According to Table 4.4, the Z_3 function created groups with a more balanced Fisher information than the Z_2 function. This function also provided higher Fisher information given to each learner than the Z_1 function. However, the overall accuracy obtained by the Z_3 function was not better than that of the Z_1 function. The values of standard deviation in Table 4.2 show that the Z_1 function tends to form groups that maximize the Fisher information given to each learner as much as possible with the smallest standard deviation. This result suggests that the optimization of groups considering the Fisher information given to each learner is crucial to improve the accuracy.

It is also worth noting that the Z_1 function, which maximizes the lower bound of the Fisher information given to each learner, does not guarantee to maximize the average value of the Fisher information of each learner although such cases were not confirmed in this experiment.

The results explained above reveal that it is difficult to improve the accuracy of ability assessment considerably if peer assessment is conducted within each group only. Because in that case, accurate peer-raters with high Fisher information can be assigned to evaluate a limit of peer-learners in a group only.



(a) $J = 15$, $G = 3$ (b) $J = 30$, $G = 5$

Figure 4.1 Fisher information for each learner in groups created by the proposed method.

To demonstrate this, we calculated the Fisher information given to each learner in the groups by using the proposed method with the Z_1 function for two cases $J = 15$, $G = 3$, and $J = 30$, $G = 5$. Figure 4.1 shows the results. In Figure 4.1, the horizontal axis denotes the ability level θ of learners. The vertical axis indicates the Fisher information $I_i(\theta_j)$. Each data point represents the Fisher information given to individual learner. The symbols of the data points denote groups to which each learner belongs. From Figure 4.1, it can be confirmed that the proposed method could not provide high Fisher information to all learners.

4.4 Summary

This chapter proposed a new group optimization method based on the IRT models that incorporate rater characteristics to maximize the accuracy of ability assessment. Concretely, the group optimization problem was formulated as an integer programming problem that maximizes the lower bound of the Fisher information given to each learner.

The experimental results using simulated data showed that the proposed method does not sufficiently improve the accuracy of peer assessment compared to the random group formation. Several alternative objective functions also showed the same tendency. These results reveal that, when peer assessment is conducted within each group only, it is difficult to improve the accuracy of ability assessment sufficiently. This result is consistent with the findings reported in Huxham and Land (2000), Pang et al. (2015), and van der Laan Smith and Spindle (2007).

Chapter 5

External Rater Selection using Item Response Theory

5.1 Introduction

As presented in Chapter 4, the proposed group optimization method based on IRT models could not sufficiently improve the accuracy of ability assessment.

To overcome that limitation, this study introduces the concept of external raters, who are peer-learners assigned to the other groups. This study also proposes an external rater selection method based on IRT that assigns a few appropriate external raters who provide higher Fisher information to each learner given the groups formed by using the proposed group optimization method. The accuracy of ability assessment is expected to be improved if each learner is additionally assessed by appropriate external raters with high Fisher information.

The following section firstly defines the concept of external raters for peer assessment when it is conducted within groups.

5.2 External rater of peer assessment conducted within each group

Definition 5.2.1. Given the groups $\boldsymbol{X}_i$ that have been formed for assignment i, a set of selectable external raters of a learner j who belongs to group g on assignment i is a set of all peer-raters assigned to group $g' \in \boldsymbol{G} \backslash g$.

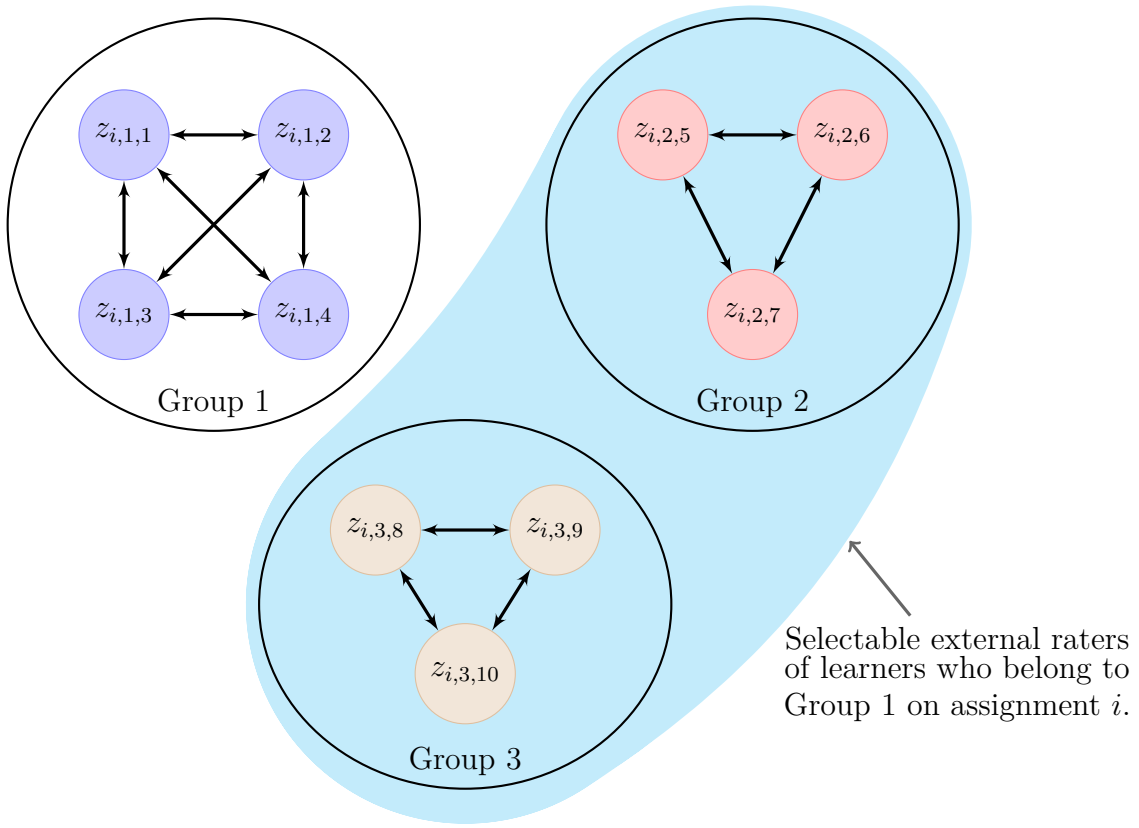


Figure 5.1 An example of selectable external raters for peer assessment. Each node $z_{i,g,j}$ presents a learner j assigned to group g on assignment i.

Let $\boldsymbol{SER}_{ij}$ be the set of selectable external raters of learner j on assignment i given the groups $\boldsymbol{X}_i$. Then, the $\boldsymbol{SER}_{ij}$ can be formulated as follows.

$$\boldsymbol{SER}_{ij} = \{r \mid r \in \boldsymbol{J},\ \sum_{g \in \boldsymbol{G}} x_{igjr} = 0\}. \tag{5.1}$$

Figure 5.1 depicts an example of selectable external raters of learners who belong to a specific group given the group construction for arbitrary assignment i. In this example, we assume that there are ten learners ($J = 10$) and these learners are divided into three groups ($G = 3$). Thus, two groups consist of three members, and one group consists of four members. In Figure 5.1, $z_{igj} = \sum_{r \in \boldsymbol{J}} x_{igjr}$ denotes learner j assigned to group g for assignment i. In this case, the selectable external raters of learners who belong to Group 1 are all learners in Group 2 and 3 as depicted in Figure 5.1. Similarly, the selectable external raters of learners who belong to Group 2 are all learners in Group 1 and 3, and the selectable external raters of learners who belong to Group 3 are all learners in Group 1 and 2, respectively.

The next section proposes a method to select a few appropriate external raters for each learner from these sets of selectable external raters.

5.3 External Rater Selection based on IRT

This section formulates the external rater selection method base on IRT models incorporating rater characteristic parameters and integer programming.

The external rater selection method assigns a few appropriate external raters to maximize the Fisher information given to each learner. Concretely, this study formulates the external rater selection problem as an integer programming problem that maximizes the lower bound of the Fisher information given by external raters to each learner. As presented in Section 5.2, given the groups $\boldsymbol{X}_i$ to which learners have been already assigned, the external rater selection method to select appropriate external raters for learner $j \in \boldsymbol{J}$ from the selectable set $\boldsymbol{SER}_{ij}$ on assignment $i \in \boldsymbol{N}$ is formulated as the following optimization problem.

$$\begin{aligned}
&\text{maximize :} && y'_i && && (5.2)\\
&\text{subject to} \\
& && \sum_{r \in \boldsymbol{SER}_{ij}} I_{ir}(\theta_j) w_{ijr} \geq y'_i, && \forall j, && (5.3)\\
& && \sum_{r \in \boldsymbol{SER}_{ij}} w_{ijr} = n^e, && \forall j, && (5.4)\\
& && \sum_{j \in \boldsymbol{J}} w_{ijr} \leq n^J, && \forall r, && (5.5)\\
& && w_{ijj} = 0, && \forall j, && (5.6)\\
& && w_{ijr} \in \{0, 1\}, && \forall j, r. && (5.7)
\end{aligned}$$

In this formulated problem, w_{ijr} is a decision variable and satisfies

$$w_{ijr} = \begin{cases} 1, & \text{if rater } r \text{ is assigned to learner } j \text{ on assignment } i, \\ 0, & \text{otherwise.} \end{cases}$$

Parameter n^e denotes the number of external raters assigned to each learner. Parameter n^J indicates the maximum number of learners that an external rater must further assess peer-learners. In this study, the setting values of n^e and n^J must satisfy $n^e \geq n^J$.

Constraints (5.3) indicate that the Fisher information given by external raters to each learner must be greater than or equal to the lower bound y'_i. Constraints (5.4) assure that each learner must be assessed by n^e external raters. Constraints (5.5) restrict that a rater can assess at most n^J learners in the other groups. These constraints are to avoid immensely increasing the assessment workload for learners. Finally, constraints (5.7) restrict that a learner cannot assess him/herself.

The objective function is defined as a maximin optimization model. As the results explained in Chapter 4, for the external rater selection problem, this study only considers the objective function that maximizes the lower bound of Fisher information given by assigned external raters.

By solving the integer programming problem, the proposed external rater selection method selects a few *appropriate* external raters who can assess the assigned learners with higher Fisher information for each learner. Consequently, by using the proposed method, the accuracy of ability assessment is expected to be improved considerably.

5.4 Evaluation using simulated data

5.4.1 Performance in comparison to random rater selection

By using the proposed method, each learner can be assessed by not only the peer-raters within groups but also appropriate external raters with high Fisher information. It is therefore expected that the accuracy of ability assessment is improved only by introducing a few external raters. This subsection conducts the following experiment to evaluate the performance of the proposed method.

1. For $J \in \{15, 30\}$ and $N \in \{4, 5\}$, the true model parameters were generated randomly from the prior distributions in Table 5.1.

2. For each assignment i, learners were divided into G groups using the proposed grouping method, *MxFiG*. Similar to the experiment in Section 4.3, $G \in \{3, 4, 5\}$ for $J = 15$, and $G \in \{3, 4, 5, 10\}$ for $J = 30$ were used.

3. Then, given the formed groups, $n^e \in \{1, 2, 3\}$ external raters were assigned to each learner using (1) the proposed method (designated as *MxFiE*), and (2) a random selection method (designated as *RndE*). In this experiment, $n^J \in \{3, 6, 12\}$ was chosen to evaluate its effects on the performance of those methods.

4. Given rater assignments and the true model parameters, rating data were sampled randomly following the IRT model.

Table 5.1 Prior distributions used for evaluating external rater selection methods.

$\theta_j \sim N(0.0, 1.0)$
$\log \alpha_r \sim N(0.0, 0.5)$, $\epsilon_r \sim N(0.0, 0.8)$
$\log \alpha_i \sim N(0.1, 0.4)$, $\beta_{ik} \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
$\boldsymbol{\mu} = (-2.0, -0.75, 0.75, 2.0)$
$\boldsymbol{\Sigma} = \begin{pmatrix} 0.16 & 0.10 & 0.04 & 0.04 \\ 0.10 & 0.16 & 0.10 & 0.04 \\ 0.04 & 0.10 & 0.16 & 0.10 \\ 0.04 & 0.04 & 0.10 & 0.16 \end{pmatrix}$

5. The ability of learners was estimated from the sampled rating data given the true parameters of raters and assignments using EAP method as similar to the experiment in Section 4.3.

6. The RMSE values between the estimated ability and the true ability were calculated by equation (4.11). The Fisher information given to each learner was also calculated.

7. After repeating the procedures described above 10 times, the mean values of the RMSE and Fisher information were calculated.

The values of the Fisher information and RMSE are presented in Table 5.2 and Table 5.3, respectively.

According to the results, both external rater selection methods revealed the higher Fisher information and the lower RMSE than the proposed group optimization method in all cases. It suggests that the introduction of the external raters is useful to improve the accuracy of ability assessment for peer assessment conducting within each group. Similar to the results of the grouping methods as presented in Section 4.3, the accuracy of the external rater selection methods tended to increase with the increasing the number of assignments N or with the decreasing the number of groups G.

The Fisher information of the external rater selection methods increased monotonically with the increase in the number of assigned external raters n^e. The RMSE also tended to be improved with the increase of n^e. However, it was not decreased monotonically when n^e increased as the Fisher information. When n^e increases, selecting appropriate external raters whose high Fisher information for each learner gradually becomes difficult because the number of suitable candidates is reduced. A small improvement in the Fisher information given to each learner, which was also confirmed in Section 4.3, would not help the RMSE decrease monotonically.

Table 5.2 Fisher information of grouping and external selection methods using simulated data.

			Grouping methods		External rater selection methods																	
					$n^J = 3$						$n^J = 6$						$n^J = 12$					
					$n^e = 1$		$n^e = 2$		$n^e = 3$		$n^e = 1$		$n^e = 2$		$n^e = 3$		$n^e = 1$		$n^e = 2$		$n^e = 3$	
J	N	G	RndG	MxFiG	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE
15	4	3	9.182	**9.604**	11.847	**13.842**	14.040	**15.700**	16.017	**16.756**	11.881	**14.366**	14.308	**17.184**	16.266	**19.097**	11.911	**14.448**	14.052	**17.853**	16.328	**19.701**
			(2.370)	(2.671)	(3.236)	(3.839)	(3.701)	(4.411)	(4.176)	(4.561)	(3.170)	(3.993)	(3.826)	(5.062)	(4.243)	(5.486)	(3.169)	(4.057)	(3.775)	(5.308)	(4.309)	(5.808)
		4	6.355	**6.426**	8.684	**10.633**	11.151	**12.766**	13.151	**13.612**	8.665	**11.310**	10.863	**14.298**	13.447	**16.010**	8.768	**11.434**	11.000	**15.256**	13.197	**17.079**
			(1.710)	(1.814)	(2.408)	(2.988)	(2.974)	(3.690)	(3.580)	(3.753)	(2.482)	(3.130)	(2.700)	(4.193)	(3.558)	(4.552)	(2.473)	(3.299)	(2.935)	(4.610)	(3.338)	(4.900)
		5	4.604	**4.780**	6.910	**9.109**	9.146	**11.046**	11.382	**12.036**	7.041	**9.816**	9.411	**12.631**	11.455	**14.609**	6.905	**10.175**	9.317	**14.074**	11.570	**15.799**
			(1.202)	(1.308)	(1.895)	(2.585)	(2.376)	(3.120)	(3.024)	(3.310)	(1.850)	(2.677)	(2.534)	(3.414)	(2.887)	(4.044)	(1.749)	(2.775)	(2.481)	(4.245)	(3.004)	(4.481)
	5	3	11.156	**11.671**	14.440	**16.834**	17.096	**19.109**	19.572	**20.400**	14.563	**17.459**	17.197	**20.881**	19.870	**23.192**	14.410	**17.607**	16.858	**21.590**	19.673	**23.968**
			(2.570)	(2.984)	(3.738)	(4.328)	(4.049)	(4.957)	(4.717)	(5.131)	(3.850)	(4.466)	(4.129)	(5.667)	(4.795)	(6.190)	(3.553)	(4.450)	(3.952)	(5.861)	(4.765)	(6.441)
		4	7.781	**7.826**	10.642	**12.975**	13.529	**15.528**	15.958	**16.580**	10.474	**13.789**	13.375	**17.432**	16.205	**19.547**	10.631	**13.953**	13.409	**18.648**	16.106	**20.827**
			(1.766)	(2.040)	(2.747)	(3.338)	(3.239)	(4.064)	(3.900)	(4.180)	(2.644)	(3.504)	(3.138)	(4.642)	(3.736)	(5.168)	(2.599)	(3.547)	(3.311)	(5.150)	(3.873)	(5.484)
		5	5.454	**5.801**	8.716	**11.111**	11.349	**13.435**	13.825	**14.660**	8.604	**11.953**	11.540	**15.426**	14.115	**17.775**	8.633	**12.378**	11.198	**17.126**	14.109	**19.289**
			(1.216)	(1.421)	(2.087)	(2.898)	(2.851)	(3.430)	(3.214)	(3.745)	(2.153)	(3.006)	(2.735)	(3.854)	(3.544)	(4.529)	(1.979)	(2.911)	(2.530)	(4.635)	(3.396)	(5.062)
30	4	3	15.853	**16.227**	17.945	**19.783**	19.664	**20.994**	21.519	**21.948**	17.965	**20.801**	19.789	**22.941**	21.527	**24.522**	18.001	**20.963**	19.798	**23.427**	21.656	**25.442**
			(4.448)	(4.741)	(5.214)	(5.977)	(5.676)	(6.123)	(6.305)	(6.493)	(5.232)	(6.608)	(5.743)	(7.111)	(6.272)	(7.473)	(5.123)	(6.624)	(5.646)	(7.156)	(6.421)	(7.482)
		4	11.719	**11.844**	13.567	**15.528**	15.293	**16.866**	17.023	**17.668**	13.622	**16.485**	15.422	**18.814**	17.258	**20.475**	13.674	**16.763**	15.376	**19.550**	17.071	**21.590**
			(3.490)	(3.524)	(4.024)	(4.827)	(4.500)	(5.148)	(5.000)	(5.377)	(4.086)	(5.420)	(4.567)	(5.991)	(5.283)	(6.455)	(4.126)	(5.400)	(4.489)	(6.330)	(4.895)	(6.842)
		5	8.876	**9.169**	10.954	**12.770**	12.701	**14.205**	14.411	**14.955**	10.894	**13.780**	12.700	**16.118**	14.387	**17.802**	10.962	**14.123**	12.692	**17.077**	14.411	**19.201**
			(2.595)	(2.774)	(3.377)	(4.019)	(3.813)	(4.368)	(4.328)	(4.610)	(3.220)	(4.581)	(3.749)	(5.207)	(4.226)	(5.619)	(3.224)	(4.691)	(3.808)	(5.546)	(4.220)	(6.199)
		10	3.525	**3.599**	5.409	**7.340**	7.155	**8.722**	8.810	**9.479**	5.345	**8.478**	7.062	**10.847**	8.792	**12.528**	5.379	**9.071**	7.106	**12.092**	8.752	**14.494**
			(0.976)	(1.029)	(1.586)	(2.368)	(2.088)	(2.674)	(2.573)	(2.921)	(1.588)	(3.122)	(2.014)	(3.602)	(2.544)	(4.063)	(1.540)	(3.520)	(2.097)	(4.456)	(2.306)	(5.130)
	5	3	20.325	**20.872**	23.055	**25.292**	25.284	**26.955**	27.436	**28.115**	23.094	**26.416**	25.370	**29.184**	27.605	**31.160**	23.039	**26.546**	25.453	**29.787**	27.657	**32.275**
			(5.052)	(5.345)	(5.836)	(6.623)	(6.372)	(6.869)	(6.992)	(7.225)	(6.018)	(7.240)	(6.402)	(7.777)	(6.910)	(8.187)	(5.849)	(7.240)	(6.548)	(7.896)	(7.238)	(8.252)
		4	14.874	**15.195**	17.630	**19.823**	19.697	**21.560**	21.937	**22.591**	17.490	**20.939**	19.605	**23.928**	21.948	**25.973**	17.534	**21.254**	19.664	**24.884**	21.948	**27.341**
			(3.864)	(3.934)	(4.658)	(5.297)	(5.027)	(5.686)	(5.645)	(5.936)	(4.457)	(5.923)	(5.045)	(6.513)	(5.505)	(7.057)	(4.491)	(6.010)	(5.023)	(7.021)	(5.696)	(7.694)
		5	11.325	**11.718**	13.935	**16.286**	16.311	**18.162**	18.450	**19.105**	13.961	**17.525**	16.152	**20.416**	18.443	**22.560**	13.949	**17.914**	16.229	**21.579**	18.479	**24.335**
			(2.777)	(3.066)	(3.656)	(4.438)	(4.306)	(4.784)	(4.772)	(5.112)	(3.678)	(5.029)	(4.182)	(5.667)	(4.694)	(6.050)	(3.613)	(5.252)	(4.222)	(6.114)	(4.682)	(6.851)
		10	4.527	**4.644**	7.002	**9.395**	9.161	**11.176**	11.411	**12.136**	6.898	**10.749**	9.170	**13.817**	11.223	**15.909**	6.876	**11.519**	9.236	**15.341**	11.478	**18.348**
			(1.126)	(1.186)	(1.845)	(2.629)	(2.287)	(2.985)	(2.882)	(3.215)	(1.854)	(3.463)	(2.311)	(3.969)	(2.930)	(4.435)	(1.791)	(3.917)	(2.377)	(5.011)	(3.001)	(5.809)

Table 5.3 RMSE values of external selection methods using simulated data.

			Grouping methods		External rater selection methods																	
					$n^J = 3$						$n^J = 6$						$n^J = 12$					
					$n^e = 1$		$n^e = 2$		$n^e = 3$		$n^e = 1$		$n^e = 2$		$n^e = 3$		$n^e = 1$		$n^e = 2$		$n^e = 3$	
J	N	G	RndG	MxFiG	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE
15	4	3	**0.315**	0.337	0.298	**0.270**	0.268	**0.247**	0.257	**0.241**	0.282	**0.252**	0.258	**0.231**	0.255	**0.222**	0.292	**0.255**	0.275	**0.231**	0.248	**0.220**
			(0.084)	(0.054)	(0.066)	(0.045)	(0.071)	(0.045)	(0.056)	(0.043)	(0.067)	(0.040)	(0.076)	(0.045)	(0.058)	(0.043)	(0.066)	(0.040)	(0.069)	(0.031)	(0.051)	(0.036)
		4	0.399	**0.396**	0.342	**0.311**	0.306	**0.292**	**0.259**	0.271	0.350	**0.289**	0.282	**0.270**	0.256	**0.239**	0.328	**0.312**	0.310	**0.256**	0.270	**0.256**
			(0.091)	(0.094)	(0.079)	(0.061)	(0.083)	(0.066)	(0.065)	(0.049)	(0.088)	(0.062)	(0.074)	(0.066)	(0.076)	(0.038)	(0.077)	(0.060)	(0.072)	(0.039)	(0.062)	(0.044)
		5	0.466	**0.447**	0.360	**0.334**	0.328	**0.295**	0.303	**0.276**	0.373	**0.321**	0.323	**0.279**	0.294	**0.261**	0.338	**0.311**	0.315	**0.269**	0.270	**0.255**
			(0.109)	(0.090)	(0.082)	(0.073)	(0.075)	(0.076)	(0.049)	(0.059)	(0.088)	(0.070)	(0.050)	(0.066)	(0.062)	(0.050)	(0.090)	(0.047)	(0.074)	(0.062)	(0.079)	(0.048)
	5	3	**0.310**	0.313	0.275	**0.256**	0.280	**0.225**	0.237	**0.222**	0.262	**0.243**	0.236	**0.218**	0.241	**0.204**	0.284	**0.241**	0.240	**0.212**	0.237	**0.197**
			(0.080)	(0.084)	(0.060)	(0.059)	(0.067)	(0.052)	(0.053)	(0.050)	(0.099)	(0.057)	(0.079)	(0.052)	(0.072)	(0.054)	(0.080)	(0.040)	(0.058)	(0.046)	(0.070)	(0.038)
		4	**0.333**	0.356	0.306	**0.278**	0.267	**0.260**	**0.242**	**0.242**	0.304	**0.266**	0.268	**0.247**	0.241	**0.222**	0.308	**0.279**	0.263	**0.236**	0.238	**0.232**
			(0.078)	(0.099)	(0.078)	(0.073)	(0.095)	(0.071)	(0.064)	(0.043)	(0.081)	(0.064)	(0.062)	(0.061)	(0.064)	(0.061)	(0.083)	(0.068)	(0.058)	(0.057)	(0.062)	(0.064)
		5	**0.395**	0.413	0.323	**0.296**	0.292	**0.255**	0.261	**0.253**	0.339	**0.299**	0.282	**0.254**	0.265	**0.239**	0.342	**0.286**	0.301	**0.248**	0.260	**0.231**
			(0.100)	(0.094)	(0.064)	(0.080)	(0.065)	(0.066)	(0.051)	(0.055)	(0.066)	(0.078)	(0.061)	(0.051)	(0.059)	(0.054)	(0.095)	(0.051)	(0.068)	(0.076)	(0.052)	(0.058)
30	4	3	0.247	**0.227**	0.220	**0.211**	0.209	**0.200**	**0.203**	0.214	0.212	**0.211**	0.217	**0.191**	0.207	**0.197**	0.214	**0.205**	0.208	**0.196**	0.211	**0.180**
			(0.050)	(0.046)	(0.043)	(0.045)	(0.046)	(0.041)	(0.035)	(0.043)	(0.038)	(0.037)	(0.035)	(0.029)	(0.039)	(0.044)	(0.038)	(0.036)	(0.039)	(0.035)	(0.043)	(0.026)
		4	**0.284**	0.292	0.268	**0.255**	0.258	**0.244**	**0.241**	0.248	0.275	**0.251**	0.245	**0.234**	0.246	**0.216**	0.269	**0.247**	0.257	**0.219**	0.235	**0.221**
			(0.060)	(0.048)	(0.040)	(0.053)	(0.049)	(0.054)	(0.046)	(0.054)	(0.043)	(0.053)	(0.031)	(0.041)	(0.039)	(0.044)	(0.043)	(0.064)	(0.036)	(0.045)	(0.042)	(0.046)
		5	0.361	**0.336**	0.318	**0.301**	0.280	**0.269**	0.273	**0.253**	0.307	**0.267**	0.292	**0.262**	0.254	**0.247**	0.304	**0.274**	0.297	**0.256**	0.268	**0.240**
			(0.054)	(0.068)	(0.063)	(0.068)	(0.041)	(0.051)	(0.061)	(0.067)	(0.070)	(0.062)	(0.075)	(0.066)	(0.048)	(0.063)	(0.062)	(0.063)	(0.076)	(0.061)	(0.061)	(0.060)
		10	0.504	**0.466**	0.404	**0.349**	0.340	**0.320**	0.343	**0.304**	0.404	**0.336**	0.364	**0.291**	0.323	**0.298**	0.404	**0.333**	0.357	**0.289**	0.317	**0.269**
			(0.071)	(0.077)	(0.056)	(0.040)	(0.066)	(0.077)	(0.048)	(0.051)	(0.073)	(0.058)	(0.077)	(0.045)	(0.085)	(0.059)	(0.075)	(0.063)	(0.059)	(0.050)	(0.066)	(0.047)
	5	3	0.233	**0.212**	0.202	**0.195**	0.198	**0.188**	0.192	**0.187**	0.203	**0.192**	0.195	**0.176**	0.184	**0.173**	0.199	**0.194**	0.189	**0.175**	0.189	**0.165**
			(0.055)	(0.042)	(0.040)	(0.037)	(0.042)	(0.035)	(0.038)	(0.037)	(0.046)	(0.035)	(0.037)	(0.022)	(0.033)	(0.036)	(0.039)	(0.038)	(0.040)	(0.023)	(0.037)	(0.020)
		4	0.260	**0.254**	0.245	**0.223**	0.223	**0.210**	**0.213**	0.218	0.230	**0.224**	0.224	**0.203**	0.217	**0.190**	0.238	**0.223**	0.223	**0.194**	0.205	**0.194**
			(0.047)	(0.037)	(0.037)	(0.044)	(0.030)	(0.047)	(0.041)	(0.035)	(0.030)	(0.036)	(0.037)	(0.027)	(0.029)	(0.026)	(0.032)	(0.052)	(0.030)	(0.030)	(0.039)	(0.032)
		5	0.294	**0.288**	0.262	**0.260**	0.257	**0.237**	0.231	**0.227**	0.260	**0.236**	0.247	**0.229**	0.226	**0.215**	0.268	**0.248**	0.249	**0.229**	0.238	**0.211**
			(0.038)	(0.052)	(0.047)	(0.051)	(0.061)	(0.039)	(0.056)	(0.051)	(0.057)	(0.051)	(0.040)	(0.051)	(0.047)	(0.053)	(0.048)	(0.058)	(0.044)	(0.048)	(0.043)	(0.045)
		10	0.444	**0.409**	0.359	**0.297**	0.303	**0.289**	**0.271**	0.274	0.342	**0.292**	0.313	**0.257**	0.291	**0.270**	0.356	**0.280**	0.313	**0.255**	0.291	**0.232**
			(0.062)	(0.072)	(0.056)	(0.031)	(0.055)	(0.074)	(0.051)	(0.053)	(0.053)	(0.048)	(0.055)	(0.033)	(0.058)	(0.054)	(0.063)	(0.054)	(0.026)	(0.040)	(0.043)	(0.042)

Compared with the random selection method, the proposed method gave higher Fisher information in all cases. The proposed method also provided lower RMSE than the random method in all cases when $n^J = 6$ and $n^J = 12$. When $n^J = 3$, however, the proposed method insufficiently improved the accuracy compared to the random method in some cases of $n^e = 3$. In the case of $n^J = 3$ and $n^e = 3$, the improvement of the Fisher information given to each learner was small in comparison with the cases of $n^J = 3$ and $n^e \in \{1, 2\}$. This result, therefore, induced the same tendency in improving the RMSE as the proposed group optimization method. The result can be explained as follows. When n^J decreases or n^e increases, selecting appropriate external raters becomes difficult because the number of suitable candidates for each learner decreases as mentioned above. The selection particularly becomes more difficult when $n^J = n^e$. In that case, all external raters must be assigned to external learners even if some of them have low Fisher information.

From the results, it can be concluded that the proposed external rater selection method enables to improve the accuracy of peer assessment sufficiently when a large value of n^J and a small value of n^e is given.

5.4.2 Effectiveness of appropriate external rater selection

In the previous experiment, this study showed that the introduction of external raters is useful to improve the accuracy of ability assessment and the proposed external rater selection method provided higher accuracy than the proposed group optimization method. In that experiment, two factors that help to improve the accuracy include (1) the increase of assigned raters and (2) the selection of appropriate external raters whose high Fisher information for each learner. The previous experiment demonstrated the effectiveness of the increasing of assigned raters for each learner. As also explained in Section 4.3, this result is a direct consequence given in inequality (3.7) and equation (3.9). However, the effects of the selecting appropriate external raters for each learner was not examined directly. Thus, this subsection demonstrates the effectiveness of the appropriate external rater selection.

For that purpose, this study first introduces another external rater selection method that assigns appropriate external raters without increasing the number of raters assigned to each learner. Concretely, the method applies the following two steps to equalize the number of raters assigned to each learner as in the proposed group optimization method.

(i) First, assign n^e external raters selected by the proposed external rater selection method (i.e., *MxFiE* method).

(ii) Then, remove n^e internal-group raters whose the lowest Fisher information.

These two steps are applied to each learner. The explained method is called as *MxFiExRs* method. If the accuracy of the *MxFiExRs* method outperforms that of the proposed group optimization method, it can be concluded that the improvement is induced by the factor of appropriate external rater selection for each learner.

To demonstrate the effectiveness of the *MxFiExRs* method, the same simulation experiment as in Subsection 5.4.1 using the *MxFiExRs* method as the external rater selection method was conducted. The results are shown in Table 5.4 and Table 5.5. It should be noted that, for the cases of $J = 15$, $G = 5$ and $J = 30$, $G = 10$, the number of internal-group raters for each learner is two. However, when $n^e = 3$ the number of raters that must be removed is three. Thus, for those cases, the results of $n^e = 3$ were not presented.

The results show that the *MxFiExRs* method provided higher Fisher information and lower RMSE than the proposed group optimization method in all cases although the number of raters for each learner was not increased. Furthermore, the *MxFiExRs* method considerably improves the accuracy as n^J increased. This result indicates that the selection of appropriate external raters for each learner plays a significant factor in improving the accuracy of ability assessment.

According to Table 5.4, it is worth noting that the the Fisher information given by the *MxFiExRs* method was not increased monotonically with the increasing of n^e, unlike in the previous experiments. Because the *MxFiExRs* method might remove internal-group raters whose higher Fisher information than the added external raters.

Table 5.4 Fisher information given to each learner induced by *MxFiExRs* method.

				MxFiExRs								
				$n^J = 3$			$n^J = 6$			$n^J = 12$		
J	N	G	MxFiG	$n^e = 1$	$n^e = 2$	$n^3 = 3$	$n^e = 1$	$n^e = 2$	$n^3 = 3$	$n^e = 1$	$n^e = 2$	$n^3 = 3$
15	4	3	9.604	**13.061**	**13.517**	**11.921**	**13.584**	**15.000**	**14.261**	**13.666**	**15.669**	**14.866**
			(2.671)	(3.737)	(4.005)	(3.497)	(3.893)	(4.684)	(4.429)	(3.962)	(4.921)	(4.749)
		4	6.426	**9.566**	**8.976**	**7.185**	**10.242**	**10.508**	**9.584**	**10.366**	**11.466**	**10.652**
			(1.814)	(2.810)	(2.796)	(1.959)	(2.952)	(3.303)	(2.813)	(3.128)	(3.713)	(3.136)
		5	4.780	**7.842**	**6.265**	-	**8.550**	**7.850**	-	**8.909**	**9.293**	-
			(1.308)	(2.331)	(1.817)	-	(2.425)	(2.147)	-	(2.538)	(2.988)	-
	5	3	11.671	**15.867**	**16.440**	**14.488**	**16.492**	**18.213**	**17.279**	**16.640**	**18.921**	**18.055**
			(2.984)	(4.185)	(4.488)	(3.881)	(4.324)	(5.242)	(4.956)	(4.319)	(5.432)	(5.226)
		4	7.826	**11.658**	**10.877**	**8.754**	**12.472**	**12.781**	**11.721**	**12.636**	**13.996**	**13.001**
			(2.040)	(3.099)	(3.027)	(2.159)	(3.268)	(3.607)	(3.205)	(3.327)	(4.119)	(3.506)
		5	5.801	**9.560**	**7.633**	-	**10.402**	**9.624**	-	**10.828**	**11.324**	-
			(1.421)	(2.611)	(2.022)	-	(2.720)	(2.478)	-	(2.651)	(3.282)	-
30	4	3	16.227	**19.341**	**19.890**	**19.938**	**20.359**	**21.836**	**22.550**	**20.520**	**22.322**	**23.470**
			(4.741)	(5.830)	(5.760)	(5.832)	(6.463)	(6.769)	(6.897)	(6.479)	(6.816)	(6.927)
		4	11.844	**15.053**	**15.604**	**15.300**	**16.010**	**17.556**	**18.122**	**16.287**	**18.293**	**19.235**
			(3.524)	(4.701)	(4.795)	(4.727)	(5.304)	(5.669)	(5.863)	(5.284)	(6.020)	(6.315)
		5	9.166	**12.265**	**12.789**	**12.187**	**13.274**	**14.687**	**14.978**	**13.616**	**15.658**	**16.382**
			(2.772)	(3.876)	(3.966)	(3.954)	(4.445)	(4.839)	(4.949)	(4.560)	(5.216)	(5.642)
		10	3.599	**6.388**	**5.124**	-	**7.526**	**7.248**	-	**6.388**	**5.124**	-
			(1.029)	(2.144)	(1.657)	-	(2.928)	(2.641)	-	(2.144)	(1.657)	-
	5	3	20.872	**24.713**	**25.518**	**25.590**	**25.838**	**27.746**	**28.597**	**25.967**	**28.349**	**29.713**
			(5.345)	(6.446)	(6.446)	(6.547)	(7.068)	(7.386)	(7.542)	(7.069)	(7.510)	(7.638)
		4	15.195	**19.210**	**19.924**	**19.565**	**20.327**	**22.295**	**22.929**	**20.642**	**23.251**	**24.299**
			(3.934)	(5.150)	(5.280)	(5.196)	(5.793)	(6.157)	(6.407)	(5.883)	(6.693)	(7.152)
		5	11.716	**15.636**	**16.335**	**15.482**	**16.874**	**18.575**	**18.928**	**17.262**	**19.747**	**20.701**
			(3.064)	(4.273)	(4.325)	(4.282)	(4.877)	(5.273)	(5.326)	(5.107)	(5.772)	(6.282)
		10	4.643	**8.160**	**6.543**	-	**9.521**	**9.184**	-	**8.160**	**6.543**	-
			(1.186)	(2.369)	(1.817)	-	(3.248)	(2.881)	-	(2.369)	(1.817)	-

Table 5.5 Comparison of RMSE values of *MxFiExRs* method with *MxFiG* method.

				MxFiExRs								
				$n^J = 3$			$n^J = 6$			$n^J = 12$		
J	N	G	MxFiG	$n^e = 1$	$n^e = 2$	$n^3 = 3$	$n^e = 1$	$n^e = 2$	$n^3 = 3$	$n^e = 1$	$n^e = 2$	$n^3 = 3$
15	4	3	0.337	**0.276**	**0.271**	**0.282**	**0.259**	**0.243**	**0.253**	**0.261**	**0.244**	**0.248**
			(0.054)	(0.041)	(0.054)	(0.048)	(0.033)	(0.051)	(0.049)	(0.048)	(0.029)	(0.047)
		4	0.396	**0.332**	**0.375**	**0.388**	**0.309**	**0.325**	**0.304**	**0.326**	**0.303**	**0.327**
			(0.094)	(0.059)	(0.078)	(0.064)	(0.058)	(0.070)	(0.069)	(0.052)	(0.048)	(0.051)
		5	0.447	**0.359**	**0.397**	-	**0.342**	**0.355**	-	**0.336**	**0.343**	-
			(0.090)	(0.079)	(0.067)	-	(0.078)	(0.089)	-	(0.054)	(0.087)	-
	5	3	0.313	**0.257**	**0.246**	**0.253**	**0.247**	**0.230**	**0.228**	**0.241**	**0.223**	**0.226**
			(0.084)	(0.059)	(0.044)	(0.038)	(0.057)	(0.047)	(0.050)	(0.041)	(0.042)	(0.035)
		4	0.356	**0.297**	**0.340**	**0.343**	**0.281**	**0.292**	**0.282**	**0.287**	**0.275**	**0.295**
			(0.099)	(0.069)	(0.096)	(0.043)	(0.063)	(0.066)	(0.085)	(0.064)	(0.062)	(0.074)
		5	0.413	**0.316**	**0.335**	-	**0.320**	**0.327**	-	**0.303**	**0.330**	-
			(0.094)	(0.084)	(0.082)	-	(0.083)	(0.080)	-	(0.056)	(0.103)	-
30	4	3	0.227	**0.211**	**0.203**	**0.217**	**0.210**	**0.190**	**0.199**	**0.205**	**0.196**	**0.184**
			(0.046)	(0.046)	(0.039)	(0.038)	(0.039)	(0.032)	(0.043)	(0.038)	(0.032)	(0.025)
		4	0.292	**0.260**	**0.253**	**0.267**	**0.255**	**0.241**	**0.220**	**0.251**	**0.226**	**0.236**
			(0.048)	(0.057)	(0.059)	(0.052)	(0.057)	(0.047)	(0.046)	(0.067)	(0.048)	(0.051)
		5	0.333	**0.302**	**0.278**	**0.276**	**0.271**	**0.274**	**0.271**	**0.275**	**0.265**	**0.263**
			(0.071)	(0.072)	(0.054)	(0.065)	(0.063)	(0.061)	(0.074)	(0.066)	(0.063)	(0.069)
		10	0.466	**0.383**	**0.403**	-	**0.359**	**0.365**	-	**0.383**	**0.403**	-
			(0.077)	(0.037)	(0.078)	-	(0.054)	(0.052)	-	(0.037)	(0.078)	-
	5	3	0.212	**0.197**	**0.192**	**0.193**	**0.196**	**0.178**	**0.178**	**0.196**	**0.177**	**0.171**
			(0.042)	(0.036)	(0.038)	(0.044)	(0.036)	(0.026)	(0.035)	(0.038)	(0.023)	(0.018)
		4	0.254	**0.227**	**0.222**	**0.236**	**0.229**	**0.214**	**0.197**	**0.227**	**0.204**	**0.209**
			(0.037)	(0.048)	(0.046)	(0.027)	(0.039)	(0.027)	(0.025)	(0.055)	(0.028)	(0.036)
		5	0.285	**0.260**	**0.239**	**0.254**	**0.238**	**0.234**	**0.236**	**0.250**	**0.234**	**0.227**
			(0.053)	(0.055)	(0.043)	(0.041)	(0.052)	(0.046)	(0.057)	(0.062)	(0.053)	(0.050)
		10	0.413	**0.334**	**0.371**	-	**0.313**	**0.329**	-	**0.334**	**0.371**	-
			(0.073)	(0.026)	(0.073)	-	(0.049)	(0.061)	-	(0.026)	(0.073)	-

5.4.3 Performance of the proposed methods with parameter estimation

This study has proposed the group optimization method and the external rater selection method based on IRT models incorporating rater characteristic parameters. The proposed methods require the parameter values of those IRT models to calculate the Fisher information given to each learner. In the previous experiments, to demonstrate the effectiveness of the proposed methods in ideal conditions, this study used the true parameter values of the Uto and Ueno (2016) model for the calculation. However, in actual e-learning situations, the parameters of IRT models are unknown and must be estimated from data.

This subsection presents a usage to apply the proposed methods to practical e-learning situations when the parameters of IRT models are unknown. Additionally, this study presents a simulation experiment to demonstrate the effectiveness of that usage.

Usage of the proposed methods with parameter estimation

This study considers the following two assumptions to use the proposed methods in actual e-learning situations.

(i) There are at least two assignments in an e-learning course.

(ii) All the assignments have been used in the past e-learning courses, and peer assessment data corresponding to that assignments were collected.

Although the second assumption might not require satisfying in practice, it is essential to estimate assignments' parameters. The Samurai system has stored all peer assessment data of all assignments used in the past courses (Uto and Ueno, 2016). In such cases, assignments' parameters can be estimated from those data.

Given the estimated parameters of assignments, the proposed methods can be applied by using the following procedures with the first assumption mentioned above.

1. For the first assignments, which is denoted as N', peer assessment is conducted using randomly formed groups. Here, the value of N' must satisfy $N' < N$.

2. The rater parameters and learner ability are estimated from the collected peer assessment data.

3. For the remaining assignments, the proposed methods then can be used given the estimated parameters.

For this method, the effectiveness of the proposed methods depends on the parameters of rater and learner ability, which were estimated from peer assessment data obtained during the first N' assignments, because the Fisher information is calculated using them. Therefore, they should be estimated as accurately as possible. In general, the estimation accuracy can be improved by increasing data size per parameter (Bishop, 2006; Uto and Ueno, 2016), which were also confirmed in the previous experiments. In the above usage, the data size per parameter can be increased by using the following two approaches.

(i) Increasing the first assignments N'.

(ii) Decreasing the number of groups created randomly during the first N' assignments (denoted as G').

The increasing of N', however, might reduce the effectiveness of the proposed methods because the number of chances to use them decreases. On the other hand, the decreasing of G' causes to increase the assessment workload to each learner during the first N' assignments. From these points of view, in practical situations, it is essential to set the value of N' as smaller as possible and the value of G' as larger as possible so that the proposed methods work appropriately.

Evaluation of the proposed methods with parameter estimation

By using the applying method presented above, this study conducted the following simulation experiment to evaluate the effects of setting values of N' and G' on the performance of the proposed methods.

1. For $J = 30$, $N = 4$, and $K = 5$, the true parameters of the IRT model were generated randomly following the prior distributions in Table 5.1.

2. For the first $N' \in \{1, 2\}$ assignments, $G' \in \{1, 2, 3\}$ groups were created randomly. In this experiment, $G' = 1$ indicates that learners were not assigned to groups.

3. Given the created groups and the true parameters, peer assessment data were sampled randomly.

4. From the sampled data and given the true assignment parameters, the rater parameters and the initial learner abilities were estimated using the Markov chain Monte Carlo (MCMC) algorithm (Uto and Ueno, 2016). The estimation also used the prior distributions in Table 5.1.

5. For the remaining assignments, $G \in \{3, 4, 5, 10\}$ groups were formed by using the *MxFiG* and *RndG* methods. Then, given the groups created by the *MxFiG* method, $n^e \in \{1, 2, 3\}$ number of external raters were assigned to each learner using the *MxFiE* and *RndE* methods. Similar to the previous experiments, $n^J \in \{3, 6, 12\}$ was used for both *MxFiE* and *RndE* methods. The Fisher information was calculated using the estimated parameters of rater and leaner ability in Step 4, and the truth values of assignment parameters generated in Step 1.

6. Given the formed groups and assigned external raters, peer assessment data were sampled randomly from the IRT model with the true parameters.

7. The learner ability were estimated using the EAP estimation method from the sampled rating data given the estimated rater parameters and the true assignment parameters.

8. The RMSE values between the estimated ability and the true ability were calculated by equation (4.11). Additionally, the Fisher information given to each learner was also calculated.

9. After repeating the steps (1-8) above 10 times, the mean and standard deviation values of the RMSE and the Fisher information were calculated.

The values of the Fisher information are presented in Table 5.6 for $N' = 1$ and Table 5.7 for $N' = 2$. And the values of the RMSE are presented in Table 5.8 for $N' = 1$ and Table 5.9 for $N' = 2$. According to the results, a similar tendency with the results of the previous simulation experiments that used the true parameters can be confirmed.

More specifically, the results showed that

(i) The proposed group optimization method could not sufficiently improve the accuracy than the random group formation.

(ii) The introduction of external raters helps to improve the accuracy considerably compared to the proposed group optimization method.

(iii) The proposed external rater selection method could improve the accuracy more sufficiently than the random selection method when a large value of n^J or a small value of n^e is given.

The results showed that the presented usage of the proposed methods works appropriately with the settings of $N' \in \{1, 2\}$ and $G' \in \{1, 2, 3\}$. As mentioned previously, it should be chosen the smallest N' and largest that the proposed methods work appropriately. Therefore, it can be concluded that $N' = 1$ and $G' = 3$ is desirable values in this experimental setting.

Additionally, it is worth noting that the decreasing of G' has a positive effect on improving the accuracy of ability assessment. The result suggests the setting of a smaller value of G' if increasing learners' assessment workload is affordable.

Table 5.6 Fisher information of the simulation experiment with parameter estimation: $N' = 1$.

			Grouping methods		External rater selection methods																	
					$n^J = 3$						$n^J = 6$						$n^J = 12$					
					$n^e = 1$		$n^e = 2$		$n^e = 3$		$n^e = 1$		$n^e = 2$		$n^e = 3$		$n^e = 1$		$n^e = 2$		$n^e = 3$	
N'	G'	G	RndG	MxFiG	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE
1	1	3	12.271	**12.464**	13.730	**15.246**	15.084	**16.237**	16.366	**16.869**	13.819	**15.882**	15.049	**17.719**	16.425	**19.111**	13.802	**16.065**	15.162	**18.257**	16.572	**19.897**
			(3.880)	(4.147)	(4.532)	(5.156)	(4.844)	(5.475)	(5.316)	(5.680)	(4.558)	(5.432)	(4.931)	(6.061)	(5.404)	(6.506)	(4.568)	(5.637)	(5.003)	(6.277)	(5.445)	(6.804)
		4	8.822	**8.863**	10.198	**11.660**	11.521	**12.679**	12.866	**13.320**	10.201	**12.322**	11.590	**14.187**	12.836	**15.558**	10.151	**12.532**	11.468	**14.898**	12.833	**16.690**
			(2.828)	(2.739)	(3.157)	(3.695)	(3.539)	(4.057)	(4.014)	(4.357)	(3.133)	(4.084)	(3.622)	(4.555)	(3.992)	(5.009)	(3.126)	(4.162)	(3.503)	(4.710)	(3.996)	(5.308)
		5	6.737	**6.972**	8.355	**9.762**	9.643	**10.827**	10.992	**11.354**	8.296	**10.435**	9.729	**12.402**	10.951	**13.765**	8.268	**10.694**	9.603	**13.130**	10.923	**14.974**
			(2.186)	(2.365)	(2.820)	(3.376)	(3.288)	(3.775)	(3.681)	(3.906)	(2.798)	(3.756)	(3.216)	(4.458)	(3.540)	(4.927)	(2.832)	(3.996)	(3.183)	(4.694)	(3.749)	(5.343)
		10	2.742	**2.785**	4.126	**5.643**	5.491	**6.668**	6.809	**7.280**	4.104	**6.381**	5.566	**8.378**	6.863	**9.754**	4.213	**6.812**	5.361	**9.405**	6.785	**11.285**
			(0.965)	(0.937)	(1.422)	(2.049)	(1.896)	(2.344)	(2.287)	(2.636)	(1.305)	(2.346)	(1.888)	(3.093)	(2.322)	(3.534)	(1.305)	(2.563)	(1.800)	(3.572)	(2.242)	(4.251)
	2	3	11.694	**11.897**	13.148	**14.344**	14.431	**15.386**	15.761	**16.079**	13.155	**14.859**	14.532	**16.657**	15.833	**17.990**	13.213	**14.991**	14.521	**17.150**	15.754	**18.821**
			(3.952)	(4.022)	(4.447)	(4.850)	(4.869)	(5.251)	(5.349)	(5.539)	(4.387)	(5.070)	(4.888)	(5.613)	(5.227)	(6.093)	(4.484)	(5.159)	(4.883)	(5.751)	(5.330)	(6.300)
		4	8.131	**8.256**	9.504	**10.637**	10.697	**11.630**	11.911	**12.251**	9.475	**11.321**	10.850	**12.889**	11.971	**14.189**	9.438	**11.540**	10.740	**13.690**	12.115	**15.250**
			(2.474)	(2.648)	(3.039)	(3.391)	(3.429)	(3.807)	(3.767)	(4.010)	(2.964)	(3.499)	(3.501)	(4.043)	(3.777)	(4.583)	(2.993)	(3.433)	(3.429)	(4.053)	(3.859)	(4.558)
		5	6.441	**6.559**	7.869	**9.034**	9.127	**10.041**	10.381	**10.721**	7.767	**9.651**	9.222	**11.382**	10.249	**12.687**	7.788	**9.854**	9.047	**12.021**	10.352	**13.677**
			(1.891)	(1.968)	(2.428)	(2.744)	(2.730)	(3.048)	(3.075)	(3.313)	(2.295)	(2.962)	(2.850)	(3.483)	(2.985)	(3.857)	(2.239)	(2.999)	(2.693)	(3.432)	(3.013)	(3.945)
		10	2.583	**2.635**	3.932	**5.120**	5.200	**6.196**	6.480	**6.847**	3.865	**5.674**	5.267	**7.499**	6.462	**8.854**	3.929	**5.933**	5.212	**8.249**	6.508	**10.159**
			(0.751)	(0.824)	(1.217)	(1.643)	(1.567)	(1.986)	(1.989)	(2.241)	(1.132)	(1.872)	(1.752)	(2.379)	(1.890)	(2.873)	(1.184)	(2.014)	(1.546)	(2.617)	(2.030)	(3.195)
	3	3	11.294	**11.365**	12.643	**13.619**	13.865	**14.609**	15.027	**15.292**	12.640	**14.177**	13.787	**15.778**	15.148	**16.987**	12.614	**14.305**	13.806	**16.366**	15.111	**17.917**
			(3.044)	(3.119)	(3.442)	(3.638)	(3.735)	(3.900)	(4.064)	(4.198)	(3.381)	(3.845)	(3.763)	(4.144)	(4.122)	(4.425)	(3.460)	(3.986)	(3.754)	(4.340)	(4.065)	(4.612)
		4	8.707	**8.839**	10.151	**11.203**	11.485	**12.404**	12.738	**13.125**	10.182	**11.623**	11.535	**13.476**	12.754	**14.882**	10.130	**11.744**	11.455	**14.118**	12.841	**15.752**
			(3.485)	(3.743)	(4.333)	(4.641)	(4.863)	(5.280)	(5.276)	(5.660)	(4.215)	(4.717)	(4.708)	(5.468)	(5.252)	(6.154)	(4.221)	(4.703)	(4.739)	(5.554)	(5.387)	(6.238)
		5	6.382	**6.514**	7.762	**8.873**	9.028	**9.974**	10.305	**10.620**	7.833	**9.339**	9.067	**11.075**	10.280	**12.460**	7.789	**9.519**	9.111	**11.736**	10.295	**13.416**
			(1.962)	(2.022)	(2.412)	(2.778)	(2.739)	(3.188)	(3.210)	(3.400)	(2.477)	(3.007)	(2.761)	(3.518)	(3.153)	(4.020)	(2.486)	(3.076)	(2.821)	(3.699)	(3.107)	(4.120)
		10	2.591	**2.624**	3.952	**5.125**	5.195	**6.197**	6.439	**6.814**	3.937	**5.549**	5.269	**7.446**	6.464	**8.852**	3.896	**5.744**	5.109	**8.125**	6.491	**10.004**
			(0.896)	(0.930)	(1.397)	(2.036)	(1.759)	(2.362)	(2.218)	(2.542)	(1.429)	(2.202)	(1.878)	(3.033)	(2.183)	(3.555)	(1.327)	(2.169)	(1.734)	(3.124)	(2.235)	(3.831)

Table 5.7 Fisher information of the simulation experiment with parameter estimation: $N' = 2$.

			Grouping methods		External rater selection methods																	
					$n^J = 3$						$n^J = 6$						$n^J = 12$					
					$n^e = 1$		$n^e = 2$		$n^e = 3$		$n^e = 1$		$n^e = 2$		$n^e = 3$		$n^e = 1$		$n^e = 2$		$n^e = 3$	
N'	G'	G	RndG	MxFiG	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE
2	1	3	7.700	**8.032**	8.885	**9.788**	9.755	**10.513**	10.526	**10.869**	8.894	**10.133**	9.792	**11.423**	10.601	**12.223**	8.890	**10.283**	9.697	**11.672**	10.636	**12.763**
			(2.630)	(2.923)	(3.205)	(3.555)	(3.551)	(3.912)	(3.772)	(4.065)	(3.214)	(3.521)	(3.533)	(4.141)	(3.772)	(4.408)	(3.207)	(3.612)	(3.540)	(4.063)	(3.857)	(4.455)
		4	5.635	**5.834**	6.693	**7.662**	7.539	**8.346**	8.411	**8.734**	6.709	**8.100**	7.554	**9.331**	8.399	**10.187**	6.739	**8.269**	7.566	**9.759**	8.452	**10.888**
			(1.959)	(2.104)	(2.364)	(2.802)	(2.585)	(3.140)	(3.018)	(3.278)	(2.423)	(2.860)	(2.756)	(3.450)	(2.917)	(3.762)	(2.424)	(2.913)	(2.680)	(3.352)	(3.074)	(3.811)
		5	4.304	**4.417**	5.267	**6.250**	6.158	**6.928**	6.962	**7.307**	5.226	**6.625**	6.137	**7.983**	6.933	**8.849**	5.320	**6.858**	6.097	**8.369**	6.999	**9.545**
			(1.481)	(1.562)	(1.886)	(2.225)	(2.143)	(2.563)	(2.421)	(2.709)	(1.849)	(2.216)	(2.149)	(2.946)	(2.332)	(3.263)	(1.894)	(2.354)	(2.107)	(2.884)	(2.500)	(3.401)
		10	1.709	1.706	2.608	**3.495**	3.458	**4.208**	4.227	**4.611**	2.549	**4.079**	3.493	**5.257**	4.186	**6.164**	2.593	**4.360**	3.470	**5.898**	4.260	**7.212**
			(0.596)	(0.577)	(0.938)	(1.180)	(1.200)	(1.519)	(1.456)	(1.753)	(0.889)	(1.395)	(1.190)	(1.831)	(1.475)	(2.223)	(0.916)	(1.627)	(1.300)	(1.874)	(1.437)	(2.530)
	2	3	7.857	**7.900**	8.808	**9.709**	9.605	**10.377**	10.491	**10.738**	8.808	**10.208**	9.645	**11.347**	10.574	**12.237**	8.806	**10.350**	9.544	**11.694**	10.495	**12.803**
			(2.565)	(2.586)	(2.867)	(3.194)	(3.128)	(3.475)	(3.430)	(3.616)	(2.861)	(3.352)	(3.189)	(3.761)	(3.434)	(4.155)	(2.889)	(3.379)	(3.057)	(3.714)	(3.391)	(4.072)
		4	5.645	**5.744**	6.641	**7.500**	7.490	**8.175**	8.397	**8.563**	6.627	**8.002**	7.505	**9.151**	8.360	**10.073**	6.676	**8.206**	7.526	**9.710**	8.340	**10.706**
			(1.935)	(1.997)	(2.266)	(2.575)	(2.584)	(2.868)	(2.986)	(3.053)	(2.315)	(2.682)	(2.617)	(3.111)	(2.956)	(3.494)	(2.382)	(2.746)	(2.621)	(3.060)	(2.953)	(3.386)
		5	4.175	**4.335**	5.177	**6.017**	5.997	**6.686**	6.798	**7.077**	5.151	**6.399**	5.941	**7.612**	6.787	**8.442**	5.123	**6.566**	5.994	**8.067**	6.894	**9.092**
			(1.526)	(1.602)	(1.892)	(2.213)	(2.214)	(2.501)	(2.451)	(2.695)	(1.828)	(2.279)	(2.122)	(2.815)	(2.442)	(3.189)	(1.865)	(2.231)	(2.191)	(2.867)	(2.564)	(3.244)
		10	1.688	**1.710**	2.622	**3.415**	3.406	**4.112**	4.243	**4.486**	2.516	**3.907**	3.397	**5.079**	4.315	**5.965**	2.620	**4.092**	3.335	**5.596**	4.232	**6.717**
			(0.617)	(0.636)	(0.989)	(1.223)	(1.194)	(1.541)	(1.572)	(1.752)	(0.916)	(1.332)	(1.252)	(1.806)	(1.541)	(2.207)	(0.994)	(1.338)	(1.214)	(1.850)	(1.566)	(2.302)
	3	3	7.786	**7.992**	8.846	**9.648**	9.688	**10.386**	10.506	**10.786**	8.822	**9.871**	9.692	**11.158**	10.516	**12.021**	8.811	**9.910**	9.629	**11.299**	10.536	**12.369**
			(3.085)	(3.288)	(3.604)	(3.908)	(3.961)	(4.248)	(4.267)	(4.454)	(3.665)	(3.958)	(4.023)	(4.471)	(4.304)	(4.898)	(3.581)	(3.929)	(3.869)	(4.473)	(4.222)	(4.867)
		4	5.447	**5.495**	6.303	**7.060**	7.176	**7.709**	7.906	**8.145**	6.310	**7.324**	7.103	**8.518**	7.982	**9.313**	6.266	**7.428**	7.195	**8.857**	7.883	**9.888**
			(1.913)	(1.956)	(2.262)	(2.452)	(2.510)	(2.718)	(2.785)	(2.966)	(2.213)	(2.447)	(2.454)	(2.895)	(2.801)	(3.211)	(2.199)	(2.449)	(2.505)	(2.849)	(2.719)	(3.095)
		5	3.971	**4.135**	4.945	**5.713**	5.721	**6.362**	6.525	**6.746**	4.963	**6.054**	5.763	**7.166**	6.514	**8.021**	4.989	**6.249**	5.778	**7.621**	6.523	**8.794**
			(1.463)	(1.623)	(1.928)	(2.192)	(2.189)	(2.503)	(2.505)	(2.695)	(1.915)	(2.283)	(2.201)	(2.699)	(2.490)	(3.076)	(1.953)	(2.405)	(2.305)	(2.675)	(2.489)	(3.211)
		10	1.676	**1.713**	2.486	**3.346**	3.384	**4.046**	4.126	**4.427**	2.506	**3.735**	3.394	**4.886**	4.186	**5.761**	2.593	**3.819**	3.328	**5.397**	4.207	**6.444**
			(0.612)	(0.675)	(0.927)	(1.287)	(1.294)	(1.628)	(1.555)	(1.801)	(0.961)	(1.446)	(1.335)	(1.854)	(1.557)	(2.286)	(1.059)	(1.532)	(1.230)	(2.017)	(1.559)	(2.291)

Table 5.8 RMSE values of the simulation experiment with parameter estimation: $N' = 1$.

			Grouping methods		External rater selection methods																	
					$n^J = 3$						$n^J = 6$						$n^J = 12$					
					$n^e = 1$		$n^e = 2$		$n^e = 3$		$n^e = 1$		$n^e = 2$		$n^e = 3$		$n^e = 1$		$n^e = 2$		$n^e = 3$	
N'	G'	G	RndG	MxFiG	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE
1	1	3	0.328	**0.322**	0.319	**0.290**	0.294	**0.287**	**0.283**	0.287	0.305	**0.296**	0.304	**0.284**	0.273	**0.269**	0.308	**0.295**	0.292	**0.280**	0.286	**0.272**
			(0.084)	(0.062)	(0.064)	(0.057)	(0.047)	(0.053)	(0.058)	(0.055)	(0.053)	(0.053)	(0.055)	(0.063)	(0.054)	(0.048)	(0.066)	(0.057)	(0.061)	(0.047)	(0.053)	(0.045)
		4	0.363	**0.356**	0.337	**0.319**	0.320	**0.313**	0.312	**0.308**	0.332	**0.331**	0.328	**0.292**	0.318	**0.282**	0.332	**0.331**	0.330	**0.297**	0.312	**0.278**
			(0.068)	(0.065)	(0.063)	(0.060)	(0.045)	(0.054)	(0.058)	(0.055)	(0.054)	(0.057)	(0.072)	(0.047)	(0.052)	(0.050)	(0.065)	(0.051)	(0.058)	(0.054)	(0.062)	(0.052)
		5	0.418	**0.400**	0.371	**0.355**	**0.350**	0.351	**0.330**	0.333	0.384	**0.347**	0.350	**0.328**	0.339	**0.310**	0.369	**0.353**	0.360	**0.324**	0.347	**0.309**
			(0.064)	(0.056)	(0.062)	(0.046)	(0.053)	(0.051)	(0.045)	(0.047)	(0.044)	(0.058)	(0.059)	(0.056)	(0.050)	(0.036)	(0.051)	(0.068)	(0.045)	(0.057)	(0.053)	(0.058)
		10	0.628	**0.589**	0.525	**0.467**	0.475	**0.398**	0.425	**0.412**	0.489	**0.423**	0.429	**0.396**	0.413	**0.360**	0.514	**0.426**	0.429	**0.380**	0.433	**0.329**
			(0.078)	(0.082)	(0.055)	(0.067)	(0.071)	(0.052)	(0.074)	(0.086)	(0.065)	(0.048)	(0.052)	(0.065)	(0.075)	(0.040)	(0.055)	(0.068)	(0.095)	(0.075)	(0.064)	(0.059)
	2	3	**0.323**	0.335	**0.317**	0.321	0.311	**0.310**	**0.311**	0.316	0.322	**0.311**	0.318	**0.308**	0.301	**0.290**	0.333	**0.305**	0.301	**0.295**	0.304	**0.283**
			(0.068)	(0.064)	(0.058)	(0.077)	(0.075)	(0.079)	(0.073)	(0.072)	(0.068)	(0.061)	(0.070)	(0.077)	(0.058)	(0.063)	(0.066)	(0.072)	(0.065)	(0.071)	(0.049)	(0.064)
		4	**0.390**	0.402	0.389	**0.377**	0.363	**0.342**	**0.343**	0.355	0.394	**0.369**	0.347	**0.344**	0.348	**0.324**	0.385	**0.357**	0.349	**0.328**	0.356	**0.312**
			(0.055)	(0.082)	(0.072)	(0.061)	(0.077)	(0.073)	(0.063)	(0.070)	(0.080)	(0.065)	(0.081)	(0.062)	(0.067)	(0.071)	(0.077)	(0.056)	(0.077)	(0.057)	(0.083)	(0.066)
		5	0.427	**0.415**	0.400	**0.379**	0.361	**0.358**	**0.358**	0.362	0.396	**0.354**	0.356	**0.347**	0.354	**0.319**	0.390	**0.366**	0.365	**0.329**	0.368	**0.325**
			(0.094)	(0.089)	(0.085)	(0.082)	(0.067)	(0.078)	(0.081)	(0.071)	(0.095)	(0.091)	(0.070)	(0.080)	(0.063)	(0.081)	(0.081)	(0.079)	(0.064)	(0.073)	(0.070)	(0.072)
		10	**0.598**	0.641	0.565	**0.482**	0.498	**0.435**	0.449	**0.441**	0.567	**0.469**	0.499	**0.425**	0.462	**0.408**	0.531	**0.448**	0.491	**0.432**	0.441	**0.378**
			(0.086)	(0.069)	(0.080)	(0.082)	(0.073)	(0.076)	(0.088)	(0.066)	(0.083)	(0.084)	(0.072)	(0.069)	(0.078)	(0.089)	(0.090)	(0.069)	(0.078)	(0.082)	(0.081)	(0.080)
	3	3	**0.352**	0.353	0.334	**0.317**	**0.326**	0.336	0.319	**0.316**	0.342	**0.320**	0.323	**0.315**	0.312	**0.306**	0.343	**0.314**	0.338	**0.302**	0.324	**0.305**
			(0.076)	(0.081)	(0.070)	(0.058)	(0.067)	(0.074)	(0.065)	(0.070)	(0.075)	(0.059)	(0.069)	(0.067)	(0.065)	(0.060)	(0.070)	(0.061)	(0.073)	(0.055)	(0.070)	(0.060)
		4	**0.385**	0.406	0.378	**0.355**	0.378	**0.357**	0.354	**0.352**	0.397	**0.361**	0.372	**0.337**	0.349	**0.334**	0.387	**0.368**	0.357	**0.335**	0.347	**0.316**
			(0.066)	(0.067)	(0.054)	(0.055)	(0.064)	(0.061)	(0.050)	(0.058)	(0.058)	(0.050)	(0.048)	(0.050)	(0.058)	(0.051)	(0.059)	(0.062)	(0.063)	(0.052)	(0.047)	(0.057)
		5	0.455	**0.426**	0.390	**0.382**	0.395	**0.373**	0.359	**0.348**	0.402	**0.388**	0.378	**0.363**	0.353	**0.332**	0.401	**0.385**	0.380	**0.352**	0.367	**0.338**
			(0.077)	(0.099)	(0.102)	(0.085)	(0.077)	(0.058)	(0.073)	(0.064)	(0.076)	(0.089)	(0.077)	(0.058)	(0.079)	(0.052)	(0.072)	(0.070)	(0.071)	(0.060)	(0.059)	(0.075)
		10	0.627	**0.600**	0.537	**0.521**	0.479	**0.457**	**0.439**	0.449	0.526	**0.499**	0.488	**0.472**	0.471	**0.450**	0.521	**0.492**	0.507	**0.478**	0.461	**0.419**
			(0.084)	(0.084)	(0.084)	(0.087)	(0.044)	(0.073)	(0.076)	(0.069)	(0.061)	(0.097)	(0.058)	(0.066)	(0.047)	(0.061)	(0.066)	(0.052)	(0.078)	(0.060)	(0.043)	(0.074)

Table 5.9 RMSE values of the simulation experiment with parameter estimation: $N' = 2$.

			Grouping methods		External rater selection methods																	
					$n^J = 3$						$n^J = 6$						$n^J = 12$					
					$n^e = 1$		$n^e = 2$		$n^e = 3$		$n^e = 1$		$n^e = 2$		$n^e = 3$		$n^e = 1$		$n^e = 2$		$n^e = 3$	
N'	G'	G	RndG	MxFiG	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE
2	1	3	**0.375**	0.380	0.370	**0.363**	0.356	**0.333**	0.348	**0.337**	0.362	**0.340**	0.346	**0.313**	0.332	**0.317**	0.369	**0.341**	0.360	**0.324**	0.331	**0.320**
			(0.046)	(0.079)	(0.070)	(0.099)	(0.088)	(0.083)	(0.076)	(0.096)	(0.070)	(0.083)	(0.083)	(0.091)	(0.086)	(0.087)	(0.072)	(0.093)	(0.067)	(0.089)	(0.086)	(0.090)
		4	0.427	**0.425**	0.397	**0.381**	0.392	**0.380**	0.361	**0.358**	0.415	**0.377**	0.395	**0.384**	0.377	**0.351**	0.400	**0.381**	0.364	**0.361**	0.377	**0.358**
			(0.087)	(0.080)	(0.083)	(0.078)	(0.087)	(0.091)	(0.093)	(0.096)	(0.084)	(0.093)	(0.070)	(0.105)	(0.093)	(0.095)	(0.080)	(0.099)	(0.086)	(0.095)	(0.089)	(0.094)
		5	**0.481**	0.491	0.441	**0.427**	0.440	**0.393**	0.395	**0.383**	0.459	**0.436**	0.419	**0.371**	0.410	**0.361**	0.459	**0.413**	0.436	**0.379**	0.386	**0.359**
			(0.087)	(0.113)	(0.099)	(0.095)	(0.111)	(0.102)	(0.089)	(0.092)	(0.116)	(0.113)	(0.098)	(0.095)	(0.083)	(0.099)	(0.122)	(0.098)	(0.091)	(0.091)	(0.096)	(0.102)
		10	0.636	**0.627**	0.545	**0.508**	**0.475**	0.478	0.473	**0.439**	0.567	**0.498**	0.497	**0.453**	0.490	**0.411**	0.517	**0.476**	0.486	**0.428**	0.471	**0.392**
			(0.108)	(0.129)	(0.115)	(0.132)	(0.112)	(0.077)	(0.105)	(0.114)	(0.118)	(0.090)	(0.094)	(0.076)	(0.085)	(0.082)	(0.108)	(0.112)	(0.097)	(0.071)	(0.082)	(0.060)
	2	3	0.402	**0.384**	0.366	**0.360**	**0.342**	0.345	**0.332**	**0.332**	0.372	**0.350**	0.350	**0.341**	0.338	**0.328**	0.362	**0.336**	0.341	**0.320**	0.345	**0.322**
			(0.098)	(0.081)	(0.074)	(0.075)	(0.076)	(0.082)	(0.086)	(0.082)	(0.083)	(0.077)	(0.070)	(0.077)	(0.083)	(0.072)	(0.086)	(0.073)	(0.066)	(0.069)	(0.079)	(0.072)
		4	**0.431**	0.449	0.423	**0.382**	0.407	**0.382**	0.397	**0.393**	0.433	**0.400**	0.394	**0.367**	0.375	**0.361**	0.424	**0.381**	0.413	**0.351**	0.396	**0.335**
			(0.130)	(0.084)	(0.078)	(0.080)	(0.087)	(0.081)	(0.092)	(0.099)	(0.086)	(0.079)	(0.089)	(0.099)	(0.058)	(0.096)	(0.080)	(0.066)	(0.078)	(0.082)	(0.080)	(0.077)
		5	**0.470**	0.481	0.454	**0.419**	0.427	**0.413**	**0.393**	0.396	0.459	**0.399**	0.417	**0.388**	0.382	**0.377**	0.451	**0.417**	0.426	**0.377**	0.405	**0.352**
			(0.078)	(0.108)	(0.135)	(0.089)	(0.110)	(0.082)	(0.087)	(0.094)	(0.093)	(0.115)	(0.091)	(0.069)	(0.096)	(0.072)	(0.121)	(0.089)	(0.078)	(0.075)	(0.104)	(0.061)
		10	0.647	**0.638**	0.557	**0.526**	**0.522**	0.534	**0.466**	0.501	0.567	**0.519**	0.526	**0.477**	0.485	**0.419**	0.575	**0.513**	0.545	**0.435**	0.500	**0.422**
			(0.133)	(0.117)	(0.127)	(0.097)	(0.107)	(0.108)	(0.107)	(0.119)	(0.109)	(0.110)	(0.123)	(0.063)	(0.105)	(0.076)	(0.119)	(0.135)	(0.087)	(0.073)	(0.135)	(0.076)
	3	3	0.413	**0.394**	0.376	**0.356**	**0.345**	0.363	**0.339**	0.357	0.379	**0.353**	0.369	**0.337**	0.351	**0.330**	0.388	**0.347**	0.353	**0.348**	**0.348**	**0.348**
			(0.075)	(0.087)	(0.082)	(0.087)	(0.078)	(0.094)	(0.075)	(0.099)	(0.091)	(0.084)	(0.088)	(0.082)	(0.076)	(0.084)	(0.090)	(0.099)	(0.083)	(0.101)	(0.092)	(0.097)
		4	**0.444**	0.451	0.425	**0.395**	0.427	**0.391**	0.383	**0.369**	0.431	**0.403**	0.408	**0.361**	0.378	**0.367**	0.419	**0.406**	0.404	**0.386**	0.380	**0.363**
			(0.115)	(0.110)	(0.102)	(0.102)	(0.116)	(0.088)	(0.103)	(0.091)	(0.093)	(0.092)	(0.095)	(0.086)	(0.086)	(0.081)	(0.097)	(0.088)	(0.102)	(0.095)	(0.107)	(0.100)
		5	0.502	**0.476**	0.450	**0.417**	**0.406**	**0.406**	0.423	**0.394**	0.453	**0.421**	0.417	**0.395**	0.406	**0.361**	0.454	**0.417**	0.431	**0.376**	0.388	**0.353**
			(0.092)	(0.085)	(0.077)	(0.084)	(0.083)	(0.096)	(0.112)	(0.120)	(0.092)	(0.071)	(0.100)	(0.069)	(0.073)	(0.066)	(0.080)	(0.093)	(0.086)	(0.082)	(0.068)	(0.080)
		10	0.651	**0.645**	0.587	**0.543**	0.536	**0.495**	**0.456**	0.477	0.574	**0.519**	0.518	**0.469**	0.484	**0.409**	0.584	**0.517**	0.520	**0.461**	0.505	**0.406**
			(0.075)	(0.119)	(0.078)	(0.077)	(0.119)	(0.097)	(0.095)	(0.108)	(0.091)	(0.076)	(0.144)	(0.100)	(0.110)	(0.084)	(0.104)	(0.062)	(0.106)	(0.067)	(0.114)	(0.079)

5.5 Evaluation using actual peer assessment data

The previous sections presented the simulation experiments using peer assessment data sampled from the IRT model to evaluate the performance of the proposed methods. The data were generated randomly by using the Monte Carlo simulation method (Spall, 2005). Therefore, it can be seen that those data were sampled under ideal conditions without any noisy effect.

In practical e-learning situations, however, actual peer assessment data might not fit any particular IRT model. This section, therefore, presents a simulation experiment using actual peer assessment data to evaluate the effectiveness of the proposed methods.

5.5.1 Data collection

The actual peer assessment data were collected using the following procedures.

1. 34 university students were collected to take part in the experiment as the learners. They were composed of 19 undergraduate, 13 master course, and two doctor course students. These students were majoring in various science fields, such as statistics, materials, chemistry, mechanics, robotics, and information science.

2. The learners were asked to complete four essay writing assignments which were used in the national assessment of educational progress (NAEP) (Persky et al., 2003; Salahu-Din et al., 2008). They were not required to have any expert knowledge or specific prior knowledge before completing these assignments.

3. After all participants completed writing essays, they were asked to evaluate all outcomes of the other learners. The assessments were conducted using a rubric which we created based on the assessment criteria for grade 12 NAEP writing (Salahu-Din et al., 2008). The assessment rubric consists of five rating categories with corresponding scoring criteria.

5.5.2 Experiment settings

Using the peer assessment data, this study conducted the following experiment, which is similar to that in Subsection 5.4.3.

As explained in Subsection 5.4.3, the proposed methods work appropriately even when the first assignments $N' = 1$. Thus, this experiment evaluates the performance of the proposed methods in the case $N' = 1$ only. The experiment procedures are as follows.

1. All parameters in the IRT model were estimated from complete assessment data using the MCMC algorithm with the prior distributions presented in Table 5.1. The estimated assignment parameters are shown in Table 5.10.

2. For the first assignment, groups were randomly generated with $G' \in \{1, 2, 3\}$.

3. The peer assessment data u_{1jr} of the first assignment were set to missing data if learner j and rater r were not in the same group.

4. Using the assessment data for the first assignment, the rater parameters and learner ability were estimated given assignment parameters obtained in Step 1.

5. For the remaining assignments $i \in \{2, \ldots, 4\}$, $G \in \{3, 4, 5, 10\}$ groups were formed by *MxFiG* and *RndG* methods. Then, given the groups formed by the *MxFiG* method, $n^e \in \{1, 2, 3\}$ external raters were assigned to each learner by *RndE*, *MxFiE*, and *MxFiExRs* methods. Similar to the experiment in Subsection 5.4.3, $n^J \in \{3, 6, 12\}$ was used.

6. Given formed groups and external raters, peer assessment data u_{ijr} were set to missing data if learner j and rater r were not in the same group and rater r was not an external rater of learner j.

7. The abilities of learners were estimated using the data, given the rater parameters and assignment parameters estimated.

8. The RMSE values between the abilities estimated in the Step 1 and those values estimated by Step 7. Concretely, the RMSE values were calculated as

$$\text{RMSE} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(\hat{\theta}_j - \hat{\theta}_j^{Step1})^2}, \tag{5.8}$$

   with $\hat{\theta}_j^{Step1}$ and $\hat{\theta}_j$ respectively are the estimated ability of learner j in Step 1 and Step 7.

   The Fisher information given to each learner was also calculated.

9. After repeating 10 times for the procedures (2-8) above, the mean and standard deviation values of the RMSE and Fisher information were calculated.

Table 5.10 Estimated assignment parameters.

	$\hat{\alpha}_i$	$\hat{\beta}_{i1}$	$\hat{\beta}_{i2}$	$\hat{\beta}_{i3}$	$\hat{\beta}_{i4}$
Assignment 1	1.179	−3.077	−1.240	0.268	1.873
Assignment 2	1.121	−3.352	−1.259	0.278	2.037
Assignment 3	1.140	−3.726	−1.613	0.033	1.790
Assignment 4	0.812	−3.581	−1.318	0.422	2.377

5.5.3 Experiment results

Table 5.11 and Table 5.12 present the Fisher information given to each learner and the RMSE. The results show similar tendencies to those obtained in the previous simulation experiments.

According to Table 5.11, the improvement in the Fisher information given by the proposed group optimization method (i.e, *MxFiG* method) was not significant compared to the random formation. As a result, the RMSE of the proposed *MxFiG* method was not sufficiently improved. It demonstrated that it is difficult for the proposed group optimization method to improve the accuracy of ability assessment considerably.

On the other hand, the RMSE in Table 5.12 shows that the introduction of external raters is useful in improving the accuracy of peer assessment. From Table 5.11 and Table 5.12, the external rater selection methods provided higher Fisher information given to each learner and lower RMSE than grouping methods. The improvement in the accuracy becomes significant when n^J is large and n^e is small.

Compared to the random selection method, the proposed external rater selection method provided the higher accuracy in all cases when $n^J = 6$ and $n^J = 12$. In particular, the RMSE of the proposed method with n^e external raters were almost equivalent to those of the random method with $n^e + 1$ external raters. When $n^J = 3$, as explained in the previous experiments, the proposed method insufficiently improved the accuracy compared to the random method in some cases of $n^e = 3$. Furthermore, as explained in Subsection 5.4.3, the proposed *MxFiE* method provided better accuracy when the value of G' was small.

The RMSE values of the *MxFiExRs* method are presented in Table 5.13. The result shows that, without increasing the number of raters assigned to each learner, the *MxFiExRs* method outperformed the *MxFiG* method in all cases. Furthermore, as an example, Figure 5.2 depicts the Fisher information given to each learner in the cases of $G' = 1, G \in \{3, 5\}$, $G' = 2, G = 3$, and $G' = 3, G = 5$. In Figure 5.2, the horizontal axis denotes each learner, and the vertical axis presents the Fisher information given

Table 5.11 Fisher information of the experiment using real data.

		Grouping methods		External rater selection methods																	
				$n^J = 3$						$n^J = 6$						$n^J = 12$					
				$n^e = 1$		$n^e = 2$		$n^e = 3$		$n^e = 1$		$n^e = 2$		$n^e = 3$		$n^e = 1$		$n^e = 2$		$n^e = 3$	
G'	G	RndG	MxFiG	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE
1	3	15.373	**15.399**	16.904	**17.927**	18.373	**19.146**	19.774	**19.991**	16.882	**18.475**	18.362	**20.466**	19.809	**22.001**	16.953	**18.596**	18.432	**21.349**	19.823	**23.153**
		(0.103)	(0.122)	(0.117)	(0.153)	(0.161)	(0.149)	(0.164)	(0.158)	(0.147)	(0.214)	(0.204)	(0.235)	(0.307)	(0.212)	(0.190)	(0.242)	(0.140)	(0.351)	(0.168)	(0.425)
	4	11.178	**11.205**	12.652	**13.735**	14.170	**14.974**	15.597	**15.816**	12.719	**14.307**	14.156	**16.313**	15.627	**17.870**	12.728	**14.480**	14.147	**17.162**	15.748	**19.065**
		(0.126)	(0.112)	(0.137)	(0.143)	(0.197)	(0.180)	(0.157)	(0.164)	(0.100)	(0.163)	(0.158)	(0.175)	(0.156)	(0.195)	(0.211)	(0.166)	(0.184)	(0.216)	(0.249)	(0.285)
	5	8.645	**8.699**	10.151	**11.247**	11.693	**12.493**	13.114	**13.310**	10.210	**11.807**	11.723	**13.850**	13.131	**15.401**	10.181	**11.985**	11.643	**14.801**	13.150	**16.735**
		(0.073)	(0.076)	(0.102)	(0.093)	(0.111)	(0.126)	(0.096)	(0.095)	(0.077)	(0.088)	(0.169)	(0.098)	(0.197)	(0.104)	(0.140)	(0.133)	(0.084)	(0.149)	(0.141)	(0.248)
	10	3.694	3.519	4.952	**6.093**	6.466	**7.336**	7.943	**8.133**	4.993	**6.688**	6.533	**8.725**	7.938	**10.277**	5.022	**6.871**	6.489	**9.807**	7.966	**11.773**
		(0.034)	(0.049)	(0.126)	(0.079)	(0.077)	(0.097)	(0.070)	(0.090)	(0.068)	(0.102)	(0.139)	(0.112)	(0.125)	(0.135)	(0.128)	(0.122)	(0.176)	(0.203)	(0.104)	(0.235)
2	3	15.682	**15.744**	17.302	**18.315**	18.719	**19.580**	20.211	**20.441**	17.269	**18.785**	18.711	**20.823**	20.272	**22.368**	17.241	**18.837**	18.735	**21.464**	20.293	**23.413**
		(1.112)	(1.172)	(1.269)	(1.398)	(1.448)	(1.493)	(1.485)	(1.539)	(1.276)	(1.407)	(1.382)	(1.568)	(1.497)	(1.731)	(1.268)	(1.397)	(1.279)	(1.514)	(1.546)	(1.506)
	4	11.431	**11.516**	13.041	**14.146**	14.494	**15.400**	15.993	**16.235**	13.027	**14.657**	14.512	**16.782**	15.992	**18.367**	13.031	**14.754**	14.546	**17.534**	16.046	**19.560**
		(0.870)	(0.910)	(0.979)	(1.163)	(1.155)	(1.296)	(1.247)	(1.324)	(1.094)	(1.297)	(1.124)	(1.465)	(1.244)	(1.623)	(1.006)	(1.314)	(1.152)	(1.582)	(1.261)	(1.734)
	5	8.552	**8.620**	10.140	**11.109**	11.610	**12.383**	12.977	**13.165**	10.112	**11.618**	11.545	**13.549**	13.021	**15.097**	10.118	**11.810**	11.527	**14.275**	12.973	**16.216**
		(0.549)	(0.605)	(0.748)	(0.847)	(0.823)	(0.951)	(0.898)	(0.962)	(0.686)	(0.930)	(0.748)	(1.069)	(0.859)	(1.189)	(0.706)	(1.069)	(0.791)	(1.146)	(0.929)	(1.252)
	10	3.707	3.616	5.153	**6.300**	6.617	**7.576**	8.126	**8.355**	5.146	**6.905**	6.651	**9.009**	8.121	**10.465**	5.126	**7.204**	6.636	**9.969**	8.112	**12.135**
		(0.253)	(0.250)	(0.346)	(0.574)	(0.459)	(0.608)	(0.537)	(0.616)	(0.388)	(0.719)	(0.433)	(0.899)	(0.564)	(1.000)	(0.350)	(0.891)	(0.492)	(1.095)	(0.430)	(1.315)
3	3	16.558	**16.720**	18.322	**19.452**	19.924	**20.863**	21.461	**21.748**	18.347	**19.931**	20.011	**22.124**	21.527	**23.857**	18.329	**20.126**	19.941	**22.900**	21.488	**25.003**
		(1.986)	(2.173)	(2.357)	(2.656)	(2.561)	(2.865)	(2.722)	(2.913)	(2.401)	(2.760)	(2.561)	(3.180)	(2.694)	(3.422)	(2.404)	(2.796)	(2.559)	(3.347)	(2.726)	(3.653)
	4	11.511	**11.566**	13.097	**14.064**	14.609	**15.428**	16.084	**16.308**	13.099	**14.472**	14.606	**16.500**	16.123	**18.151**	13.096	**14.554**	14.630	**17.088**	16.165	**19.100**
		(1.014)	(1.111)	(1.266)	(1.380)	(1.350)	(1.571)	(1.501)	(1.566)	(1.227)	(1.377)	(1.372)	(1.674)	(1.528)	(1.881)	(1.246)	(1.381)	(1.438)	(1.715)	(1.568)	(2.081)
	5	9.062	**9.129**	10.718	**11.670**	12.255	**13.097**	13.726	**13.964**	10.655	**12.076**	12.238	**14.223**	13.786	**15.922**	10.690	**12.275**	12.216	**14.815**	13.694	**16.973**
		(0.846)	(0.898)	(1.057)	(1.145)	(1.185)	(1.333)	(1.297)	(1.373)	(1.014)	(1.175)	(1.212)	(1.384)	(1.266)	(1.615)	(1.068)	(1.264)	(1.153)	(1.378)	(1.380)	(1.628)
	10	3.895	3.700	5.222	**6.405**	6.833	**7.712**	8.351	**8.575**	5.260	**6.893**	6.810	**9.095**	8.393	**10.677**	5.228	**7.122**	6.795	**9.904**	8.360	**12.213**
		(0.283)	(0.286)	(0.384)	(0.466)	(0.517)	(0.581)	(0.609)	(0.679)	(0.398)	(0.544)	(0.528)	(0.678)	(0.600)	(0.781)	(0.426)	(0.579)	(0.507)	(0.761)	(0.558)	(0.886)

Table 5.12 RMSE values of the the experiment using real data.

		Grouping methods		External rater selection methods																	
				$n^J = 3$						$n^J = 6$						$n^J = 12$					
				$n^e = 1$		$n^e = 2$		$n^e = 3$		$n^e = 1$		$n^e = 2$		$n^e = 3$		$n^e = 1$		$n^e = 2$		$n^e = 3$	
G'	G	RndG	MxFiG	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE	RndE	MxFiE
1	3	**0.257**	0.266	0.257	**0.244**	0.250	**0.231**	**0.235**	0.242	0.257	**0.227**	0.246	**0.234**	0.235	**0.228**	0.255	**0.241**	0.244	**0.216**	0.244	**0.214**
		(0.024)	(0.028)	(0.030)	(0.024)	(0.028)	(0.020)	(0.026)	(0.013)	(0.023)	(0.022)	(0.027)	(0.014)	(0.026)	(0.014)	(0.028)	(0.025)	(0.026)	(0.022)	(0.017)	(0.021)
	4	**0.286**	0.327	0.309	**0.291**	0.286	**0.277**	0.281	**0.262**	0.310	**0.275**	0.279	**0.274**	0.282	**0.257**	0.299	**0.286**	0.295	**0.250**	0.284	**0.241**
		(0.034)	(0.028)	(0.029)	(0.033)	(0.034)	(0.025)	(0.022)	(0.022)	(0.042)	(0.025)	(0.025)	(0.025)	(0.026)	(0.025)	(0.023)	(0.021)	(0.031)	(0.012)	(0.018)	(0.021)
	5	**0.320**	0.345	0.325	**0.311**	0.309	**0.287**	0.298	**0.297**	0.327	**0.288**	0.317	**0.282**	0.291	**0.278**	0.327	**0.296**	0.294	**0.274**	0.288	**0.252**
		(0.029)	(0.024)	(0.030)	(0.025)	(0.034)	(0.025)	(0.015)	(0.040)	(0.030)	(0.036)	(0.032)	(0.024)	(0.017)	(0.019)	(0.018)	(0.022)	(0.016)	(0.018)	(0.026)	(0.021)
	10	**0.426**	0.445	0.423	**0.379**	0.362	**0.343**	0.368	**0.323**	0.401	**0.373**	0.380	**0.340**	0.332	**0.329**	0.404	**0.372**	0.365	**0.312**	0.331	**0.290**
		(0.054)	(0.062)	(0.049)	(0.032)	(0.043)	(0.047)	(0.032)	(0.027)	(0.052)	(0.053)	(0.044)	(0.032)	(0.043)	(0.025)	(0.035)	(0.048)	(0.056)	(0.028)	(0.041)	(0.033)
2	3	**0.276**	0.281	0.274	**0.267**	0.266	**0.258**	**0.245**	0.253	0.274	**0.258**	0.260	**0.238**	0.264	**0.244**	0.275	**0.257**	0.263	**0.244**	0.259	**0.234**
		(0.056)	(0.037)	(0.037)	(0.045)	(0.042)	(0.039)	(0.051)	(0.034)	(0.027)	(0.044)	(0.044)	(0.033)	(0.047)	(0.043)	(0.038)	(0.040)	(0.042)	(0.044)	(0.046)	(0.056)
	4	**0.306**	0.314	0.304	**0.277**	0.289	**0.281**	0.278	**0.265**	0.303	**0.288**	0.288	**0.275**	0.273	**0.259**	0.304	**0.284**	0.294	**0.255**	0.276	**0.244**
		(0.040)	(0.036)	(0.025)	(0.044)	(0.044)	(0.038)	(0.026)	(0.037)	(0.028)	(0.044)	(0.028)	(0.040)	(0.038)	(0.041)	(0.031)	(0.047)	(0.032)	(0.037)	(0.033)	(0.041)
	5	0.348	**0.339**	0.318	**0.311**	0.297	**0.295**	0.295	**0.280**	0.316	**0.304**	0.302	**0.284**	0.289	**0.272**	0.324	**0.294**	0.300	**0.285**	0.306	**0.271**
		(0.067)	(0.011)	(0.021)	(0.032)	(0.026)	(0.027)	(0.025)	(0.021)	(0.017)	(0.023)	(0.028)	(0.037)	(0.026)	(0.028)	(0.018)	(0.026)	(0.024)	(0.035)	(0.024)	(0.027)
	10	**0.462**	0.470	0.398	**0.390**	**0.372**	0.380	0.352	**0.341**	0.437	**0.367**	0.391	**0.360**	0.348	**0.331**	0.432	**0.389**	0.390	**0.350**	0.363	**0.296**
		(0.059)	(0.067)	(0.049)	(0.031)	(0.036)	(0.045)	(0.043)	(0.056)	(0.079)	(0.038)	(0.045)	(0.039)	(0.045)	(0.056)	(0.061)	(0.053)	(0.062)	(0.039)	(0.025)	(0.039)
3	3	0.292	**0.291**	0.285	**0.267**	**0.271**	0.273	**0.257**	0.283	0.280	**0.269**	0.274	**0.261**	0.262	**0.251**	0.279	**0.264**	0.275	**0.256**	0.263	**0.242**
		(0.039)	(0.040)	(0.040)	(0.039)	(0.044)	(0.042)	(0.044)	(0.055)	(0.038)	(0.046)	(0.030)	(0.028)	(0.038)	(0.027)	(0.035)	(0.049)	(0.040)	(0.041)	(0.044)	(0.043)
	4	0.338	**0.332**	0.319	**0.308**	0.305	**0.301**	0.306	**0.303**	0.315	**0.279**	0.292	**0.276**	0.296	**0.273**	0.307	**0.290**	0.313	**0.288**	0.300	**0.271**
		(0.043)	(0.036)	(0.035)	(0.042)	(0.034)	(0.038)	(0.036)	(0.039)	(0.032)	(0.030)	(0.038)	(0.028)	(0.034)	(0.042)	(0.033)	(0.030)	(0.039)	(0.023)	(0.029)	(0.022)
	5	0.366	**0.348**	0.337	**0.317**	0.321	**0.316**	0.314	**0.300**	0.350	**0.316**	0.336	**0.310**	0.312	**0.294**	0.346	**0.322**	0.320	**0.306**	0.317	**0.279**
		(0.049)	(0.051)	(0.061)	(0.045)	(0.049)	(0.055)	(0.049)	(0.045)	(0.059)	(0.045)	(0.056)	(0.036)	(0.056)	(0.047)	(0.036)	(0.058)	(0.051)	(0.068)	(0.053)	(0.043)
	10	**0.464**	0.494	0.431	**0.416**	0.418	**0.389**	0.382	**0.364**	0.448	**0.428**	0.388	**0.367**	0.368	**0.340**	0.421	**0.407**	0.410	**0.376**	0.372	**0.324**
		(0.034)	(0.064)	(0.050)	(0.057)	(0.042)	(0.035)	(0.034)	(0.042)	(0.048)	(0.050)	(0.037)	(0.043)	(0.046)	(0.027)	(0.038)	(0.055)	(0.030)	(0.055)	(0.036)	(0.025)

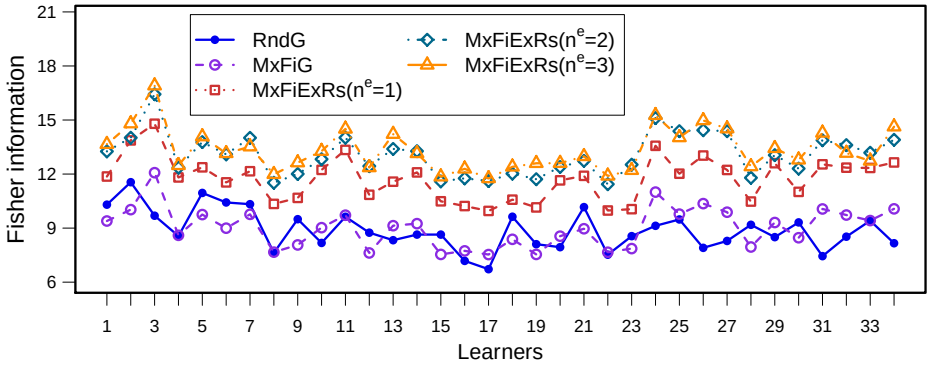
Table 5.13 RMSE values of the *MxFiExRs* using real data.

			MxFiExRs								
			$n^J = 3$			$n^J = 6$			$n^J = 12$		
G'	G	MxFiG	$n^e = 1$	$n^e = 2$	$n^3 = 3$	$n^e = 1$	$n^e = 2$	$n^3 = 3$	$n^e = 1$	$n^e = 2$	$n^3 = 3$
1	3	0.271	**0.251**	**0.256**	**0.256**	**0.248**	**0.247**	**0.242**	**0.250**	**0.244**	**0.233**
		(0.032)	(0.030)	(0.022)	(0.021)	(0.028)	(0.024)	(0.020)	(0.032)	(0.027)	(0.027)
	4	0.296	**0.267**	**0.272**	**0.266**	**0.252**	**0.267**	**0.261**	**0.276**	**0.259**	**0.271**
		(0.028)	(0.027)	(0.020)	(0.020)	(0.033)	(0.030)	(0.016)	(0.022)	(0.020)	(0.025)
	5	0.329	**0.297**	**0.305**	**0.303**	**0.294**	**0.293**	**0.292**	**0.304**	**0.274**	**0.293**
		(0.035)	(0.047)	(0.039)	(0.039)	(0.028)	(0.040)	(0.039)	(0.033)	(0.028)	(0.024)
	10	0.435	**0.379**	0.439	**0.383**	**0.373**	**0.404**	**0.385**	**0.431**	**0.399**	**0.360**
		(0.053)	(0.047)	(0.058)	(0.045)	(0.031)	(0.054)	(0.038)	(0.045)	(0.024)	(0.030)
2	3	0.284	**0.254**	**0.253**	**0.275**	**0.257**	**0.258**	**0.261**	**0.253**	**0.238**	**0.245**
		(0.033)	(0.025)	(0.025)	(0.036)	(0.025)	(0.027)	(0.036)	(0.022)	(0.025)	(0.035)
	4	0.338	**0.315**	**0.313**	**0.320**	**0.306**	**0.304**	**0.319**	**0.297**	**0.292**	**0.291**
		(0.051)	(0.056)	(0.060)	(0.031)	(0.043)	(0.050)	(0.047)	(0.046)	(0.048)	(0.043)
	5	0.340	**0.308**	**0.304**	**0.320**	**0.301**	**0.288**	**0.303**	**0.313**	**0.283**	**0.290**
		(0.029)	(0.027)	(0.041)	(0.036)	(0.024)	(0.029)	(0.028)	(0.025)	(0.021)	(0.026)
	10	0.469	**0.415**	**0.436**	**0.420**	**0.411**	**0.399**	**0.389**	**0.401**	**0.405**	**0.382**
		(0.051)	(0.045)	(0.050)	(0.056)	(0.054)	(0.043)	(0.044)	(0.064)	(0.041)	(0.037)
3	3	0.287	**0.267**	**0.260**	**0.256**	**0.276**	**0.267**	**0.252**	**0.285**	**0.254**	**0.266**
		(0.026)	(0.029)	(0.035)	(0.031)	(0.029)	(0.024)	(0.035)	(0.035)	(0.025)	(0.024)
	4	0.328	**0.304**	**0.293**	**0.291**	**0.310**	**0.291**	**0.289**	**0.291**	**0.288**	**0.274**
		(0.036)	(0.038)	(0.035)	(0.032)	(0.038)	(0.030)	(0.039)	(0.032)	(0.029)	(0.042)
	5	0.358	**0.335**	**0.338**	**0.337**	**0.338**	**0.340**	**0.321**	**0.324**	**0.330**	**0.333**
		(0.066)	(0.063)	(0.057)	(0.037)	(0.042)	(0.031)	(0.035)	(0.057)	(0.044)	(0.050)
	10	0.485	**0.445**	**0.471**	**0.441**	**0.435**	**0.434**	**0.437**	**0.447**	**0.435**	**0.416**
		(0.041)	(0.057)	(0.073)	(0.063)	(0.060)	(0.042)	(0.030)	(0.068)	(0.038)	(0.051)

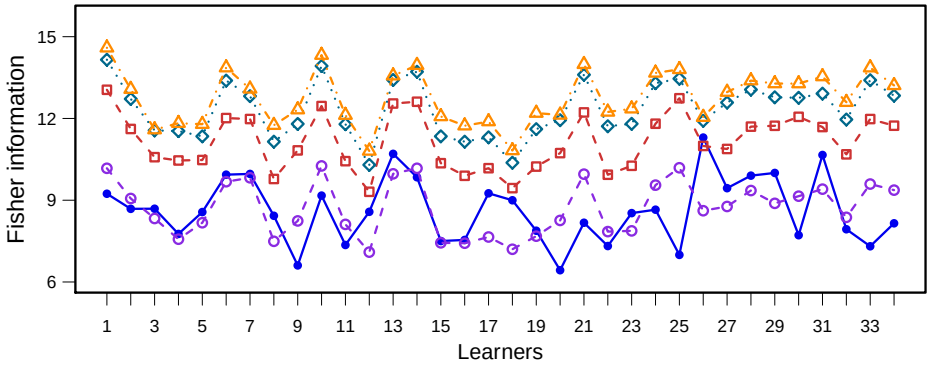
to each learner. From Figure 5.2, the Fisher information given to each learner induced by the *MxFiExRs* method is higher than those of the *MxFiG* and *RndG* methods.

From these results, it is demonstrated that the proposed external rater selection method, which enables to select a few appropriate external raters to each learner, efficiently improves the accuracy of peer assessment.

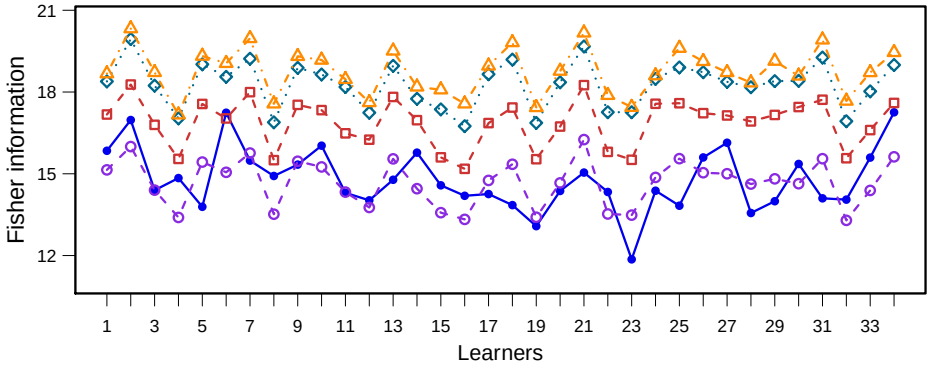
(a) $N' = 1, G' = 1, G = 3$



(b) $N' = 1, G' = 1, G = 5$



(c) $N' = 1, G' = 2, G = 3$



(d) $N' = 1, G' = 3, G = 5$

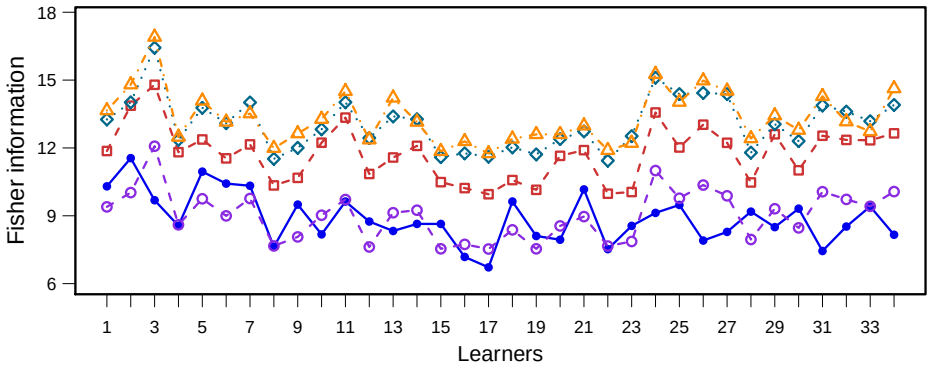


Figure 5.2 An example of the Fisher information given to each learner of actual data.

5.5.4 Example of estimated parameters and rater assignment

This subsection presents an example of the created groups and external raters assigned to each learner by the proposed methods for Assignment 2. Additionally, the estimated values of rater and ability parameters in the IRT model and examples of item characteristic curves are also presented.

Specifically, Table 5.14 presents an example in the case of $G' = 3$, $G = 5$, $n^J = 6$, and $n^e = 3$. In Table 5.14, columns named $[\hat{\alpha}_r]$ and $[\hat{\epsilon}_r]$ present the estimated consistency and severity parameters in Step 4. Column named $[\hat{\theta}_j]$ shows the estimated learner ability in Step 4. Column named [Group members] shows the group member of each learner. Column named [External raters] presents assigned external raters of each learner. And column named [Assigned external learners] shows external learners assigned to each external rater.

From Table 5.14, it can be confirmed that, the assessment consistency and assessment severity of learners are different among learners. By using IRT models incorporating rater characteristic parameters such as the model proposed by Uto and Ueno (2016), learner ability θ_j is estimated considering these rater characteristics. To illustrate these differences among experimented learners, Figure 5.3 depicts item characteristic curves of Rater 7, 22, 27 and 33 with parameters of the Assignment 2 as shown in Table 5.10. In each panel of Figure 5.3, the horizontal axis denotes ability level θ. The left vertical axis shows the response probability to each category and the right vertical axis presents the Fisher information. From Table 5.14 and Figure 5.3, rater characteristics of the given learners can be explained as follows.

(i) Rater 7 is an extremely inconsistent rater. As presented in Subsection 3.3.2, inconsistent raters assess peer-learners with low accuracy because their ratings do not reflect learner ability accurately. As a result, in Figure 5.3, the Fisher information given by Rater 7 is extremely low.

(ii) Rater 22 is highly consistent rater with averaged severity. It means that the rater can accurately assess almost peer-learners. Thus, the Fisher information given by the rater is considerably high in the overall of ability level compared to the other raters.

(iii) Rater 27 is the most lenient rater with averaged consistency. It means that the rater cannot accurately assess peer-learners with ability in the high range. Therefore, in the high range of ability level, the Fisher information given by this rater is low.

Table 5.14 Estimated parameters, group members, and assigned external raters in the experiment given $G' = 3$, $G = 5$, $n^J = 6$, and $n^e = 3$.

Learner ($r = j$)	$\hat{\alpha}_r$	$\hat{\epsilon}_r$	$\hat{\theta}_j$	Group members	External raters	Assigned external learners
1	1.000	0.000	−0.370	{2,5,14,19,22,32}	{3,8,27}	{8,12}
2	0.897	−0.143	0.887	{1,5,14,19,22,32}	{3,26,28}	{-}
3	1.599	0.762	0.880	{8,10,18,28,29}	{9,22,27}	{1,2,13,14,26,34}
4	1.265	0.136	0.492	{6,11,12,16,17,34}	{9,15,18}	{9,13,18,20,25,30}
5	0.936	0.043	0.200	{1,2,14,19,22,32}	{15,18,27}	{-}
6	1.049	−0.138	−1.135	{4,11,12,16,17,34}	{8,24,26}	{-}
7	0.556	1.207	0.042	{15,20,21,23,24,25}	{16,27,28}	{-}
8	1.890	0.558	0.604	{3,10,18,28,29}	{1,16,24}	{1,6,15,19,23,27}
9	1.271	0.920	−0.978	{13,26,27,30,31,33}	{4,12,15}	{3,4,28,29,32}
10	0.719	−0.298	1.210	{3,8,18,28,29}	{11,19,23}	{-}
11	1.535	−0.504	1.014	{4,6,12,16,17,34}	{22,26,33}	{10,15,19,22,24,33}
12	1.346	−0.271	0.717	{4,6,11,16,17,34}	{1,24,28}	{9,20,22,24,27,33}
13	1.041	1.095	−0.081	{9,26,27,30,31,33}	{3,4,24}	{-}
14	1.165	0.045	−0.600	{1,2,5,19,22,32}	{3,16,27}	{25,26,30}
15	1.932	0.221	0.539	{7,20,21,23,24,25}	{8,11,26}	{4,5,9,16,26,32}
16	1.749	0.065	−1.891	{4,6,11,12,17,34}	{15,19,22}	{7,8,14,22,24,29}
17	0.921	−0.627	0.949	{4,6,11,12,16,34}	{18,19,23}	{-}
18	1.367	0.308	0.669	{3,8,10,28,29}	{4,23,24}	{4,5,17,27,31,32}
19	1.696	0.874	−0.103	{1,2,5,14,22,32}	{8,11,26}	{10,16,17,21,33,34}
20	0.963	1.021	−0.124	{7,15,21,23,24,25}	{4,12,22}	{-}
21	0.701	0.156	0.494	{7,15,20,23,24,25}	{19,26,28}	{-}
22	2.036	0.761	−0.110	{1,2,5,14,19,32}	{11,12,16}	{3,11,16,20,25,30}
23	1.589	0.166	−1.009	{7,15,20,21,24,25}	{8,28,33}	{10,17,18,28,29,31}
24	1.802	0.197	−1.067	{7,15,20,21,23,25}	{11,12,16}	{6,8,12,13,18,28}
25	0.786	−0.181	−0.343	{7,15,20,21,23,24}	{4,14,22}	{-}
26	1.249	0.376	0.331	{9,13,27,30,31,33}	{3,14,15}	{2,6,11,15,19,21}
27	1.334	−0.641	−0.401	{9,13,26,30,31,33}	{8,12,18}	{1,3,5,7,14,34}
28	1.755	0.296	0.532	{3,8,10,18,29}	{9,23,24}	{2,7,12,21,23,31}
29	0.881	−0.130	0.330	{3,8,10,18,28}	{9,16,23}	{-}
30	1.177	1.167	−0.332	{9,13,26,27,31,33}	{4,14,22}	{-}
31	0.911	−0.262	0.152	{9,13,26,27,30,33}	{18,23,28}	{-}
32	0.872	−0.422	0.497	{1,2,5,14,19,22}	{9,15,18}	{-}
33	1.172	0.269	−0.305	{9,13,26,27,30,31}	{11,12,19}	{11,23}
34	0.815	−0.623	−0.224	{4,6,11,12,16,17}	{3,19,27}	{-}

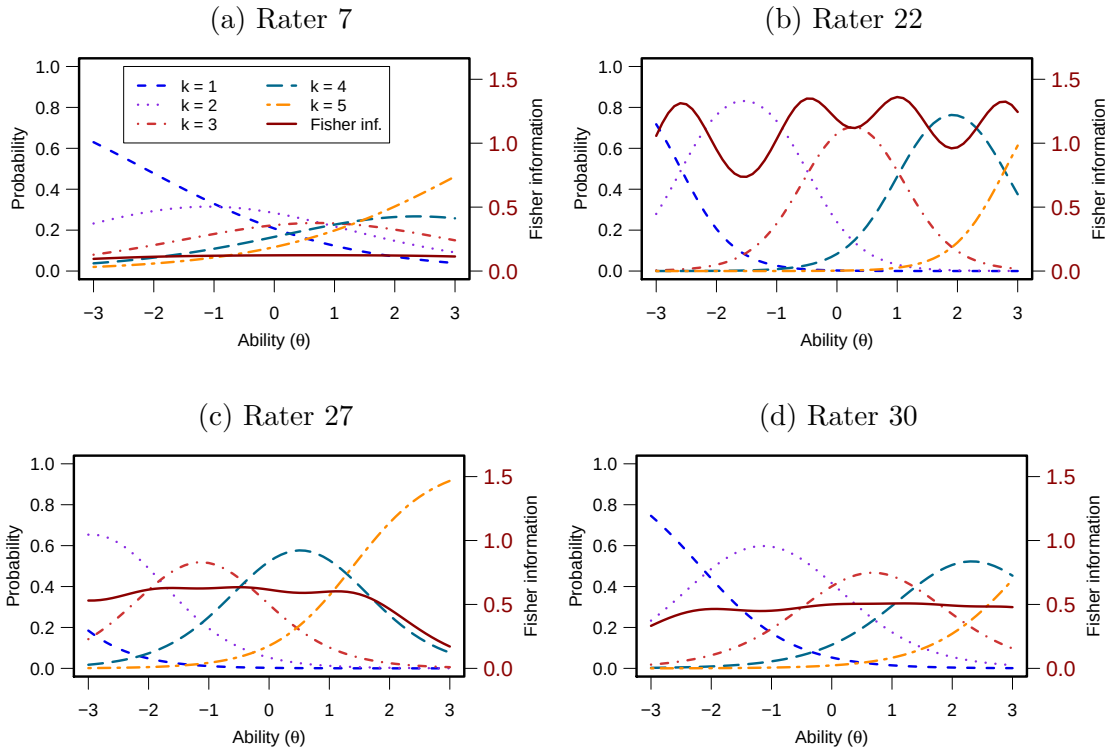


Figure 5.3 Item characteristic curves of four raters in the actual peer assessment experiment.

(iv) Rater 30 is the second most severe rater with averaged consistency. Therefore, this rater can more accurately assess peer-learners with ability in the high range compared to Rater 27.

The results from Table 5.14 reveal that the proposed group optimization method created groups that contain equivalent group members. Furthermore, the results of external raters show that different sets of external raters were assigned to each learner. In other words, the appropriate external raters of each learner are different and depend on both learner ability and rater characteristics.

Finally, the results listed in the column [Assigned external learners] of Table 5.14 reveal that a learner who cannot accurately evaluate peer-learners' outcomes would not be employed to assess external learners.

5.6 Summary

To overcome the insufficient improvement of the accuracy of grouped peer assessment, this chapter introduced the concept of external raters and proposed an external rater

selection method based on IRT models. The external rater selection problem was formulated as an integer programming problem that maximizes the lower bound of the Fisher information given by external raters to each learner.

Results of experiments using simulated data showed that the proposed method could considerably improve the accuracy of ability assessment. In particular, the results reveal that assigning appropriate peer-raters to each learner plays a significant factor in improving the accuracy of peer assessment.

This chapter also presented an application method to practical e-learning situations in which the IRT model parameters are estimated from data. Experimental results demonstrated that the proposed methods appropriately work as in ideal conditions.

Finally, this chapter presented an experiment using actual peer assessment data. Results demonstrated that it is difficult for the proposed group optimization method to improve the accuracy of ability assessment. On the other hands, the introduction of external raters helped to improve the accuracy of ability assessment sufficiently compared to the group optimization methods. Furthermore, it was confirmed that the proposed external rater selection method outperforms the random method.

Chapter 6

Conclusion

6.1 Conclusion

This study proposed group optimization methods to improve the accuracy of ability assessment when peer assessment is conducted within each group.

Chapter 2 provided a literature review on the existing group formation methods in collaborative learning. The review showed that there was no study on group optimization methods to maximize the accuracy of peer assessment conducted within each group.

In Chapter 3, this study formulated peer assessment data and then introduced an IRT model that incorporates rater characteristic parameters. This chapter also presented the Fisher information, a widely used index to evaluate the accuracy of ability assessment.

Chapter 4 proposed a group optimization method to maximize the accuracy of peer assessment conducting within each group based on IRT models incorporating rater parameters. The method was formulated as an integer programming problem that maximizes the lower bound of the Fisher information given to each learner. However, experimental results showed that the proposed method could not sufficiently improve the accuracy of ability assessment. The results revealed that it is difficult for the proposed method to improve the accuracy significantly if peer assessment is conducted within each group only.

To overcome that difficulty, in Chapter 5, this study relaxed the constraint that restricted peer assessment to be conducted within each group only by introducing the concept of external raters. This study then proposed an external rater selection method based on IRT models to assign a few appropriate external raters to each learner. The external rater selection problem was formulated as an integer programming problem

that maximizes the lower bound of the Fisher information given by external raters to each learner. Experimental results using both simulated and actual peer assessment data showed that the introduction of external raters is useful to improve the accuracy of peer assessment considerably. Furthermore, the results showed that the proposed method could significantly improve the accuracy than the random method.

Chapter 5 also presented a usage to apply the proposed methods above to actual e-learning situations, which the parameters of IRT models are unknown and must be estimated from data. Experiments using both simulated and actual data showed that the usage worked appropriately.

6.2 Future work

Several future research directions follow from this study. Firstly, this study proposed a group optimization method and an external rater selection method that maximize the lower bound of the Fisher information given to each learner. This maximin approach, however, does not guarantee the average value of the Fisher information given to each learner to be maximized. This consideration motivates a future research direction to investigate the performance of the proposed methods using a multi-objective optimization approach.

Secondly, this study formulated the optimization problems as integer programming problems based on IRT models incorporating rater parameters. To evaluate the effectiveness of the proposed methods, this study employed the IRT model proposed by Uto and Ueno (2016). Recently, other IRT models that incorporate rater characteristic parameters have also been proposed (e.g, DeCarlo, 2005; Patz et al., 2002; Ueno and Okamoto, 2008; Usami, 2010). Thus, it is valuable to analyze further the performance of the proposed methods using those similar IRT models.

Thirdly, although this study introduced a usage to apply the proposed methods to practical e-learning situations, the presented usage cannot dynamically capture changes in the values of rater parameters and learner ability throughout multiple assessments. It is preferable to an approach that enables to capture such changes adaptively. A research direction that investigates the performance of that adaptive approach to the accuracy of peer assessment might be meaningful.

Finally, this study has only focused on the group optimization methods to improve the accuracy of ability assessment when peer assessment is conducted within each group. As discussed in Chapter 2, assessment might positively influence on achievements of collaborative learning (Lan et al., 2011; Sluijsmans and Strijbos, 2010; Strijbos, 2011).

Collaborative learning in environments such as MOOCs would be benefited from highly accurate evaluation and appropriate feedback given by appropriate peer-learners. Therefore, a research direction that examines the effectiveness of the proposed methods on learning achievements provides an insightful understanding of the relationship between appropriate assessment and learning achievement.

Bibliography

Adema, J. J. (1989). Implementations of the branch-and-bound method for test construction problems. project psychometric aspects of item banking no. 46. Research Report 89-6, Department of Education, University of Twente.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4):561–573.

ArchMiller, A., Fieberg, J., Walker, J., and Holm, N. (2016). Group peer assessment for summative evaluation in a graduate-level statistics course for ecologists. *Assessment & Evaluation in Higher Education*, pages 1–13.

Baartman, L. K., Prins, F. J., Kirschner, P. A., and Van Der Vleuten, C. P. (2007). Determining the quality of competence assessment programs: A self-evaluation procedure. *Studies in Educational Evaluation*, 33(3-4):258–281.

Baker, F. B. and Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques*. New York, NY: Marcel Dekker, Inc.

Baker, K. and Powell, S. (2002). Methods for assigning students to groups: A study of alternative objective functions. *Journal of the Operational Research Society*, 53(4):397–404.

Bhalerao, A. and Ward, A. (2001). Towards electronically assisted peer assessment: a case study. *ALT-J: Research in Learning Technology*, 9(1):26–37.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Black, P. and Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1):7–74.

Bostock, S. (2000). Student peer assessment. *Learning Technology*, 5.

Bowen, S. and Martens, P. J. (2006). A model for collaborative evaluation of university-community partnerships. *Journal of Epidemiology & Community Health*, 60(10):902–907.

Brauer, S. and Schmidt, T. C. (2012). Group formation in elearning-enabled online social networks. In *Proceedings of the 15th International Conference on Interactive Collaborative Learning (ICL 2012)*, pages 1–8.

Brimberg, J., Janićijević, S., Mladenović, N., and Urošević, D. (2017). Solving the clique partitioning problem as a maximally diverse grouping problem. *Optimization Letters*, 11(6):1123–1135.

Brimberg, J., Mladenović, N., and Urošević, D. (2015). Solving the maximally diverse grouping problem by skewed general variable neighborhood search. *Information Sciences*, 295:650–675.

Brusco, M. J. and Köhn, H.-F. (2009). Clustering qualitative data based on binary equivalence relations: Neighborhood search heuristics for the clique partitioning problem. *Psychometrika*, 74(4):685–703.

Burke, B. (1998). Evaluating for a change: Reflections on participatory methodology. *New Directions for Evaluation*, 1998(80):43–56.

Capuano, N., Loia, V., and Orciuoli, F. (2017). A fuzzy group decision making model for ordinal peer assessment. *IEEE Transactions on Learning Technologies*, 10(2):247–259.

Chan, C. K. and Van Aalst, J. (2004). Learning, assessment and collaboration in computer-supported environments. In *What we know about CSCL*, pages 87–112. Boston, MA: Kluwer Academic Publishers.

Cho, K. and Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education*, 48(3):409–426.

Cho, Y., Je, S., Yoon, Y. S., Roh, H. R., Chang, C., Kang, H., and Lim, T. (2016). The effect of peer-group size on the delivery of feedback in basic life support refresher training: a cluster randomized controlled trial. *BMC Medical Education*, 16(1):167.

Christodoulopoulos, C. E. and Papanikolaou, K. A. (2007). A group formation tool in an e-learning context. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, volume 2, pages 117–123.

Conley-Tyler, M. (2005). A fundamental choice: Internal or external evaluation? *Evaluation Journal of Australasia*, 4:3–11.

Dascalu, M.-I., Bodea, C.-N., Lytras, M., De Pablos, P. O., and Burlacu, A. (2014). Improving e-learning communities through optimal composition of multidisciplinary learning groups. *Computers in Human Behavior*, 30:362–371.

Davies, P. (2007). Review in computerized peer-assessment. will it have an effect on student marking consistency? In *Proceedings of the 11th International Conference on Computer Assisted Assessment (CAA)*, pages 143–151.

DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement*, 42(1):53–76.

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

Dochy, F., Gijbels, D., and Segers, M. (2006). Learning and the emerging new assessment culture. In *Instructional Psychology: Past, Present, and Future Trends: Sixteen Essays in Honour of Eric de Corte*, pages 191–206. Oxford: Elsevier.

Dochy, F., Segers, M., and Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher education*, 24(3):331–350.

Falchikov, N. (2005). *Improving Assessment Through Student Involvement: Practical solutions for aiding learning in higher and further education*. New York, NY: RoutledgeFalmer.

Feo, T. A. and Khellaf, M. (1990). A class of bounded approximation algorithms for graph partitioning. *Networks*, 20(2):181–195.

Freeman, M. (1995). Peer assessment by groups of group work. *Assessment & Evaluation in Higher Education*, 20(3):289–300.

Frieden, B. R. (2004). *Science from Fisher information: a unification*. Cambridge University Press.

Gallego, M., Laguna, M., Martí, R., and Duarte, A. (2013). Tabu search with strategic oscillation for the maximally diverse grouping problem. *Journal of the Operational Research Society*, 64(5):724–734.

Glance, D., Forsey, M., and Riley, M. (2013). The pedagogical foundations of massive open online courses. *First Monday*, 18(5).

Gogoulou, A., Gouli, E., Boas, G., Liakou, E., and Grigoriadou, M. (2007). Forming homogeneous, heterogeneous and mixed groups of learners. In *Proceedings of the Workshop on Personalization in e-Learning Environments at Individual and Group Level at the 11th International Conference on User Modeling*, pages 33–40.

Graf, S. and Bekele, R. (2006). Forming heterogeneous groups for intelligent collaborative learning systems with ant colony optimization. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems (ITS 2006)*, pages 217–226.

Hamer, J., Ma, K. T., and Kwong, H. H. (2005). A method of automatic grade calibration in peer assessment. In *Proceedings of the 7th Australasian Conference on Computing Education*, pages 67–72.

Hübscher, R. (2010). Assigning students to groups using general and context-specific criteria. *IEEE Transactions on Learning Technologies*, 3(3):178–189.

Huxham, M. and Land, R. (2000). Assigning students in group work projects. can we do better than random? *Innovations in Education and Teaching International*, 37(1):17–22.

IBM Corp. (2015). *IBM ILOG CPLEX Optimization Studio: CPLEX User's Manual*. IBM Corporation, 12.6 edition.

Ikeda, M., Go, S., and Mizoguchi, R. (1997). Opportunistic group formation. In *Proceedings of the 8th International Conference on Artificial Intelligence in Education (AIED 1997)*, volume 97, pages 167–174.

Inaba, A., Supnithi, T., Ikeda, M., Mizoguchi, R., and Toyoda, J. (2000). How can we form effective collaborative learning groups? In *Proceedings of the 5th International Conference on Intelligent Tutoring Systems (ITS 2000)*, pages 282–291. Springer.

Jonsson, A. and Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, 2(2):130–144.

Kardan, A. A. and Sadeghi, H. (2016). An efficacious dynamic mathematical modelling approach for creation of best collaborative groups. *Mathematical and Computer Modelling of Dynamical Systems*, 22(1):39–53.

Khandaker, N. and Soh, L.-K. (2010). Improving group selection and assessment in an asynchronous collaborative writing application. *International Journal of Artificial Intelligence in Education*, 20(3):231–268.

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., and Klemmer, S. R. (2013). Peer and self assessment in massive online classes. *ACM Transactions on Computing Human Interaction*, 20(6):33:1–33:31.

Kvale, S. (2007). Contradictions of assessment for learning in institutions of higher learning. In *Rethinking Assessment in Higher Education: Learning for the longer term*, pages 57–71. New York, NY: Routledge.

Lan, C.-H., Graf, S., Lai, K. R., and Kinshuk, K. (2011). Enrichment of peer assessment with agent negotiation. *IEEE Transactions on Learning Technologies*, 4(1):35–46.

Lin, S. S., Liu, E. Z.-F., and Yuan, S.-M. (2001). Web-based peer assessment: feedback for students with various thinking-styles. *Journal of Computer Assisted Learning*, 17(4):420–432.

Lin, Y. S., Chang, Y. C., and Chu, C. P. (2016). Novel approach to facilitating tradeoff multi-objective grouping optimization. *IEEE Transactions on Learning Technologies*, 9(2):107–119.

Lin, Y.-T., Huang, Y.-M., and Cheng, S.-C. (2010). An automatic group composition system for composing collaborative learning groups using enhanced particle swarm optimization. *Computers & Education*, 55(4):1483–1493.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lu, J. and Law, N. (2012). Online peer assessment: Effects of cognitive and affective feedback. *Instructional Science*, 40(2):257–275.

Lynn Snow, A., Cook, K. F., Lin, P.-S., Morgan, R. O., and Magaziner, J. (2005). Proxies and other external raters: Methodological considerations. *Health Research and Educational Trust*, 40(5p2):1676–1693.

Mahdi, B. J. M. and Fattaneh, T. (2013). A semi-pareto optimal set based algorithm for grouping of students. In *Proceedings of the 4th IEEE International Conference on E-Learning and E-Teaching (ICELET 2013)*, pages 10–13.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174.

Matteucci, M. and Stracqualursi, L. (2006). Student assessment via graded response model. *Statistica*, 66(4):435–447.

Moccozet, L. and Tardy, C. (2015). An assessment for learning framework with peer assessment of group works. In *Proceedings of the 14th International Conference on Information Technology Based Higher Education and Training (ITHET 2015)*, pages 1–5.

Montazer, G. A. and Mohammad, S. R. (2013). E-learners grouping in uncertain environment using fuzzy art-snap-drift neural network. In *Proceedings of the 4th IEEE International Conference on E-Learning and E-Teaching (ICELET 2013)*, pages 112–116.

Montazer, G. A. and Rezaei, M. S. (2012). A new approach in e-learners grouping using hybrid clustering method. In *Proceedings of the IEEE International Conference on Education and e-Learning Innovations (ICEELI 2012)*, pages 1–5.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2):159–176.

Muraki, E., Hombo, C. M., and Lee, Y.-W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement*, 24(4):325–337.

Navrat, P. and Tvarozek, J. (2014). Online programming exercises for summative assessment in university courses. In *Proceedings of the 15th ACM International Conference on Computer Systems and Technologies*, pages 341–348.

Nevo, D. (1994). Combining internal and external evaluation: A case for school-based evaluation. *Studies in Educational Evaluation*, 20(1):87–98.

Nevo, D. (2001). School evaluation: internal or external? *Studies in Educational Evaluation*, 27(2):95–106.

Nguyen, D.-T., Uto, M., Abe, Y., and Ueno, M. (2015). Reliable peer assessment for team-project-based learning using item response theory. In *Proceedings of the 23rd International Conference on Computers in Education (ICCE 2015)*, pages 144–153.

Ounnas, A., Davis, H. C., and Millard, D. E. (2009). A framework for semantic group formation in education. *Journal of Educational Technology & Society*, 12(4):43–55.

Pang, Y., Mugno, R., Xue, X., and Wang, H. (2015). Constructing collaborative learning groups with maximum diversity requirements. In *Proceedings of the 15th IEEE International Conference on Advanced Learning Technologies (ICALT 2015)*, pages 34–38.

Pang, Y., Xiao, F., Wang, H., and Xue, X. (2014). A clustering-based grouping model for enhancing collaborative learning. In *Proceedings of the 13th IEEE International Conference on Machine Learning and Applications (ICMLA 2014)*, pages 562–567.

Papinczak, T., Young, L., and Groves, M. (2007). Peer assessment in problem-based learning: A qualitative study. *Advances in Health Sciences Education*, 12(2):169–186.

Paravati, G., Lamberti, F., Gatteschi, V., Demartini, C., and Montuschi, P. (2017). Point cloud-based automatic assessment of 3d computer animation courseworks. *IEEE Transactions on Learning Technologies*, 10(4):532–543.

Patz, R. J. and Junker, B. W. (1999). Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing Data, and Rated Responses. *Journal of Educational and Behavioral Statistics*, 24(4):342–366.

Patz, R. J., Junker, B. W., Johnson, M. S., and Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, 27(4):341–384.

Peavy, K. M., Guydish, J., Manuel, J. K., Campbell, B. K., Lisha, N., Le, T., Delucchi, K., and Garrett, S. (2014). Treatment adherence and competency ratings among therapists, supervisors, study-related raters and external raters in a clinical trial of a 12-step facilitation for stimulant users. *Journal of Substance Abuse Treatment*, 47(3):222–228.

Persky, H., Daane, M., and Jin, Y. (2003). The nation's report card: Writing 2002. Technical report, National Center for Education Statistics.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. (2013). Tuned models of peer assessment in MOOCs. In *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)*, pages 153–160.

Pollalis, Y. A. and Mavrommatis, G. (2009). Using similarity measures for collaborating groups formation: A model for distance learning environments. *European Journal of Operational Research*, 193(2):626–636.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 19(1):49–57.

Rodriguez, F. J., Lozano, M., García-Martínez, C., and GonzáLez-Barrera, J. D. (2013). An artificial bee colony algorithm for the maximally diverse grouping problem. *Information Sciences*, 230:183–196.

Rubens, N., Vilenius, M., and Okamoto, T. (2009). Automatic group formation for informal collaborative learning. In *Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 231–234.

Ryan, K. E., Chandler, M., and Samuels, M. (2007). What should school-based evaluation look like? *Studies in Educational Evaluation*, 33(3-4):197–212.

Sadeghi, H. and Kardan, A. A. (2015). A novel justice-based linear model for optimal learner group formation in computer-supported collaborative learning environments. *Computers in Human Behavior*, 48:436–447.

Sadler, P. M. and Good, E. (2006). The impact of self-and peer-grading on student learning. *Educational Assessment*, 11(1):1–31.

Salahu-Din, D., Persky, H., and Miller, J. (2008). The nation's report card: Writing 2007. Technical report, National Center for Education Statistics.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(1):1–97.

Savoia, E., Testa, M. A., Biddinger, P. D., Cadigan, R. O., Koh, H., Campbell, P., and Stoto, M. A. (2009). Assessing public health capabilities during emergency preparedness tabletop exercises: reliability and validity of a measurement tool. *Public Health Reports*, 124(1):138–148.

Seethamraju, R. and Borman, M. (2009). Influence of group formation choices on academic performance. *Assessment & Evaluation in Higher Education*, 34(1):31–40.

Shah, N. B., Bradley, J., Balakrishnan, S., Parekh, A., Ramchandran, K., and Wainwright, M. J. (2014). Some scaling laws for MOOC assessments. In *Proceedings of the KDD Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2014)*.

Shapiro, S. M., Lancee, W. J., and Richards-Bentley, C. M. (2009). Evaluation of a communication skills program for first-year medical students at the university of toronto. *BMC Medical Education*, 9(1):11.

Sitthiworachart, J. and Joy, M. (2004). Effective peer assessment for learning computer programming. In *ACM SIGCSE Bulletin*, volume 36, pages 122–126.

Sluijsmans, D. M., Moerkerke, G., van Merriënboer, J. J., and Dochy, F. J. (2001). Peer assessment in problem based learning. *Studies in Educational Evaluation*, 27(2):153–173.

Sluijsmans, D. M. and Strijbos, J.-W. (2010). Flexible peer assessment formats to acknowledge individual contributions during (web-based) collaborative learning. In *E-Collaborative Knowledge Construction: Learning from Computer-Supported and Virtual Environments*, pages 139–161. Hershey, PA: IGI Global.

Soh, L.-K., Khandaker, N., and Jiang, H. (2006). Multiagent coalition formation for computer-supported cooperative learning. In *Proceedings of the 21st National Conference on Artificial Intelligence*, volume 21, pages 1844–1851.

Soh, L.-K., Khandaker, N., and Jiang, H. (2008). I-MINDS: A multiagent system for intelligent computer-supported collaborative learning and classroom management. *International Journal of Artificial Intelligence in Education*, 18(2):119–151.

Spall, J. C. (2005). *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. Hoboken, NJ: John Wiley & Sons, Inc.

Srba, I. and Bielikova, M. (2015). Dynamic group formation as an approach to collaborative learning support. *IEEE Transactions on Learning Technologies*, 8(2):173–186.

Staubitz, T., Petrick, D., Bauer, M., Renz, J., and Meinel, C. (2016). Improving the peer assessment experience on mooc platforms. In *Proceedings of the 3rd ACM Conference on Learning@ Scale*, pages 389–398.

Strijbos, J.-W. (2011). Assessment of (computer-supported) collaborative learning. *IEEE Transactions on Learning Technologies*, 4(1):59–73.

Suen, H. K. (2014). Peer assessment for massive open online courses (MOOCs). *The International Review of Research in Open and Distributed Learning*, 15(3):312–327.

Sung, Y.-T., Chang, K.-E., Chiou, S.-K., and Hou, H.-T. (2005). The design and application of a web-based self-and peer-assessment system. *Computers & Education*, 45(2):187–202.

Tanimoto, S. L. (2007). The squeaky wheel algorithm: Automatic grouping of students for collaborative projects. In *Proceedings of the Workshop on Personalization in e-Learning Environments at Individual and Group Level at the 11th International Conference on User Modeling*, pages 79–80.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, 68(3):249–276.

Torres, R. T., Preskill, H. S., and Piontek, M. E. (1997). Communicating and reporting: Practices and concerns of internal and external evaluators. *Evaluation Practice*, 18(2):105–125.

Trahasch, S. (2004). From peer assessment towards collaborative learning. In *Proceeding of the 34th IEEE Frontiers in Education Conference (FIE 2004)*, pages F3F–16.

Ueno, M. (2004). Data mining and text mining technologies for collaborative learning in an ilms "samurai". In *Proceedings of the 4th IEEE International Conference on Advanced Learning Technologies (ICALT 2004)*, pages 1052–1053.

Ueno, M. and Okamoto, T. (2008). Item response theory for peer assessment. In *Proceedings of the 8th IEEE International Conference on Advanced Learning Technologies (ICALT 2008)*, pages 554–558.

Ueno, M., Songmuang, P., Okamoto, T., and Nagaoka, K. (2008). Item response theory with assessors' parameters of peer assessment. *IEICE Transactions on Information and Systems*, 91(2):377–388. (in Japanese).

Ueno, M. and Uto, M. (2011). Learning community using social network service. In *Proceedings of International Conference on Web Based Communities and Social Media (WBC 2011)*, pages 109–119.

Usami, S. (2010). A polytomous item response model that simultaneously considers bias factors of raters and examinees: Estimation through a Markov Chain Monte Carlo algorithm. *The Japanese Journal of Educational Psychology*, 58(2):163–175. (in Japanese).

Uto, M. and Ueno, M. (2016). Item response theory for peer assessment. *IEEE Transactions on Learning Technologies*, 9(2):157–170.

van der Laan Smith, J. and Spindle, R. M. (2007). The impact of group formation in a cooperative learning environment. *Journal of Accounting Education*, 25(4):153–167.

Van der Linden, W. J. (2006). *Linear Models for Optimal Test Design*. New York, NY: Springer Science & Business Media, Inc.

Volkov, B. B. (2011). Beyond being an evaluator: The multiplicity of roles of the internal evaluator. *New Directions for Evaluation*, 2011(132):25–42.

Volkov, B. B. and Baron, M. E. (2011). Issues in internal evaluation: Implications for practice, training, and research. *New Directions for Evaluation*, 2011(132):101–111.

Vygotsky, L. S. (1978). *Mind in society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Wang, D.-Y., Lin, S. S., and Sun, C.-T. (2007). Diana: A computer-supported heterogeneous grouping system for teachers to conduct successful small learning groups. *Computers in Human Behavior*, 23(4):1997–2010.

Wang, Z. and Yao, L. (2013). The effects of rater severity and rater distribution on examinees' ability estimation for constructed response items. *ETS Research Report Series*, 2013(2):i–22.

Weaver, R. L. and Cotrell, H. W. (1986). Peer evaluation: A case study. *Innovative Higher Education*, 11(1):25–39.

Weitz, R. and Lakshminarayanan, S. (1998). An empirical comparison of heuristic methods for creating maximally diverse groups. *Journal of the Operational Research Society*, pages 635–646.

Withey, M., Daft, R. L., and Cooper, W. H. (1983). Measures of perrow's work unit technology: An empirical assessment and a new scale. *Academy of Management Journal*, 26(1):45–63.

Wright, M. C., Segall, N., Hobbs, G., Phillips-Bute, B., Maynard, L., and Taekman, J. M. (2013). Standardized assessment for evaluation of team skills: validity and feasibility. *Simulation in Healthcare*, 8(5):292–303.

Yannibelli, V. and Amandi, A. (2011). Forming well-balanced collaborative learning teams according to the roles of their members: An evolutionary approach. In *Proceeding of the 12th IEEE International Symposium on Computational Intelligence and Informatics (CINTI 2011)*, pages 265–270.

Zakrzewska, D. (2009). Cluster analysis in personalized e-learning systems. In *Intelligent Systems for Knowledge Management*, pages 229–250. Springer-Verlag Berlin Heidelberg.

Zheng, Z. and Pinkwart, N. (2014). A discrete particle swarm optimization approach to compose heterogeneous learning groups. In *Proceedings of the 14th IEEE International Conference on Advanced Learning Technologies (ICALT 2014)*, pages 49–51.

Appendix A

List of Publications

Journal Papers

1. Nguyen, D.-T., Uto, M., and Ueno, M. (2018). Group optimization using item response theory for peer assessment. *IEICE Transactions on Information and Systems*, J101-D(2):431–445. (in Japanese).

International Conferences (Refereed)

1. Uto, M., Nguyen, D.-T., and Ueno, M. (2017). Group optimization to maximize peer assessment accuracy using item response theory. In *Proceedings of the 18th International Conference on Artificial Intelligence in Education (AIED 2017)*, pages 393–405.

2. Nguyen, D.-T., Uto, M., Abe, Y., and Ueno, M. (2015). Reliable peer assessment for team-project-based learning using item response theory. In *Proceedings of the 23rd International Conference on Computers in Education (ICCE 2015)*, pages 144–153.

Other Papers (Not Refereed)

1. Nguyen, D.-T., Uto, M., and Ueno, M. (2017). A grouping method for optimizing peer assessment accuracy. In *Proceedings of the 33rd Annual Conference of JSET*, pages 1035–1036.

2. Uto, M., Nguyen, D.-T., and Ueno, M. (2017). Automated grouping system to maximize peer assessment accuracy. In *Proceedings of the 42nd Annual Conference of JSiSE*, pages 201–202. (in Japanese).

3. Nguyen, D.-T., Uto, M., and Ueno, M. (2017). A group formation method to optimize peer assessment accuracy. In *Proceedings of the 15th Annual Conference of JART*, pages 160–163. (in Japanese).

4. Nguyen, D.-T., Uto, M., and Ueno, M. (2016). Group formation for peer assessment using item response theory. In *Proceedings of the 32nd Annual Conference of JSET*, pages 1013–1014.

5. Nguyen, D.-T., Uto, M., and Ueno, M. (2016). Rater selection method to optimize peer assessment accuracy. In *Proceedings of the 14th Annual Conference of JART*, pages 66–69. (in Japanese).