Intern. J. Comput.-Support. Collab. Learn
DOI 10.1007/s11412-016-9234-6

Quantitative approach to collaborative learning: performance prediction, individual assessment, and group composition

Ling Cen[1] **· Dymitr Ruta**[1] **· Jason Ng**[1] **· Leigh Powell**[1] **·**
Benjamin Hirsch[1]

Received: 11 February 2015 / Accepted: 25 April 2016

Abstract The benefits of collaborative learning, although widely reported, lack the quantitative rigor and detailed insight into the dynamics of interactions within the group, while individual contributions and their impacts on group members and their collaborative work remain hidden behind joint group assessment. To bridge this gap we intend to address three important aspects of collaborative learning focused on quantitative evaluation and prediction of group performance. First, we use machine learning techniques to predict group performance based on the data of member interactions and thereby identify whether, and to what extent, the group's performance is driven by specific patterns of learning and interaction. Specifically, we explore the application of Extreme Learning Machine and Classification and Regression Trees to assess the predictability of group academic performance from live interaction data. Second, we propose a comparative model to unscramble individual student performances within the group. These performance are then used further in a generative mixture model of group grading as an explicit combination of isolated individual student grade expectations and compared against the actual group performances to define what we coined as collaboration synergy - directly measururing the improvements of collaborative learning. Finally the impact of group composition of gender and skills on learning performance and collaboration synergy is evaluated. The analysis indicates a high level of predictability of group performance based solely on the style and mechanics of collaboration and quantitatively supports the claim that heterogeneous groups with the diversity of skills and genders benefit more from collaborative learning than homogeneous groups.

Keywords Collaborative learning · Performance prediction · Machine learning · Performance modeling · Group composition

✉ Ling Cen
cen.ling@kustar.ac.ae

[1] Etisalat, British Telecom Innovation Center, Khalifa University of Science, Technology and Research, Abu Dhabi, UAE

Introduction Q4

Collaborative learning (CL) refers to situations and environments in which learners engage in common tasks and each individual capitalizes on resources and skills from one another (Bruffee 1993; Dillenbourg 1999; Mitnik et al. 2009). It is based on the Q5 model that knowledge can be created within a population where members actively interact by sharing experiences and take on asymmetrical roles (Chiu 2000, 2008). Q6 Computer-supported collaborative learning (CSCL) denotes a pedagogical approach characterized by the sharing and construction of knowledge among participants using technology as their primary means of communication or as a common resource. In this approach, learning can either synchronously or asynchronously take place in online and classroom learning environments via social interaction using computers or through the Internet (Stahl et al. 2006). CSCL continues to thrive on the back of the rapid growth in cheap and powerful knowledge access technologies connecting and enabling students to carry out ever more learning, coursework and assessment tasks together (Dillenbourg 1999; Ruta et al. 2013; Hirsch et al. 2013; Dirkx and Smith 2013; Davidson and Sternberg 2003; Barkley et al. 2004, and has been widely considered as a method to improve learning performance (Zheng and Huang 2016).

In collaborative learning, however, the behavior and, thereby, the learning patterns observed are much more complex than that of individual learning. While there is a wide body of qualitative evidence reporting the benefits of collaborative learning, the thorough quantitative analysis is clearly lagging behind in the literature. This is perhaps due to difficulties with formal knowledge representation and the lack of data capturing the complete process of collaborative learning in sufficient detail. To address this, a collaborative learning environment (CLE) platform has been developed at Etisalat British Telecom Innovation Centre (EBTIC) (Ruta et al. 2013; Hirsch et al. 2013). It was trialed over one semester in the courses of the Molecular Biology Engineering and the Freshman Engineering Design at Khalifa University. During the trial, collaborative learning styles and their dynamics and outcomes were evaluated using three, group-based, formally assessed assignments.

This work is grounded in the fields of Educational Data Mining (EDM) and CSCL and builds upon prior work on collaborative learning and data-driven learning analysis, which is aimed to develop quantitative approaches to describe the characteristics of collaborative learning and assess their impact on learning performance. There are many theories on how and why group collaboration works, but most attribute it to information exchange, conflict resolution, intersubjective meaning-making, group knowledge building, and participatory models (Suthers 2006). In our research we focused on several aspects of group knowledge building and its quantitative assessment, monitoring, evaluation and prediction in order to gain more informed and measurable insight into the mechanics and quality of collaborative learning, in conjunction with its performance and key impact factors. Specifically, this work intends to address three important issues in collaborative learning focused on quantitative evaluation and prediction of group performance.

First, we explore the predictability of academic performance based on the mechanics of interactions during live collaborative learning. The aim is to predict how well the group is likely to perform given all available individual and group historical evidence as well as live interaction patterns. Predicting academic performance of students engaged in individual learning has been explored largely based on data mining and machine learning (ML) technologies in the literature, (e.g. Thai-Nghe et al. 2011a; Yadav and Pal 2012; Romero, Ventura, Espejo, & Hervs 2012). Although these models can provide an accurate prediction of learning performance for individual students, they do not account for interaction and collaboration among students within groups. It has been shown that collaboration and interaction patterns in collaborative learning can affect learning outcomes, and therefore cannot be ignored when considering impacts on collaborative learning performance (McNely et al. 2012). Prediction of group academic performance can help to evaluate and improve collaborative learning systems, identify effective grouping, design efficient interaction patterns, and help to understand what drives student academic performance in a dynamic and connected learning environment at every stage of the group exercise. For instance, prior predictions could offer recommendations as to which course, modules or specific content is suitable for the particular student or a group of students, or could aid in forming optimal group composition (i.e. the one that maximizes the expected performance). Predictions during the course could also help to identify significant deviations of the early progress from the initial expectations and identify the source of underperformance, allowing for corresponding corrective intervention. Predictions after the course, on the other hand, allow one to compare pure data driven reflection on the group performance vs the performance perceived by the teacher and hence flag any cases of significant dissonance.

A complete case study on group academic performance prediction has been carried out with the data collected in the trial of the CLE platform, which involves generation and extraction of features from the CLE group interaction data, development of machine learning models to predict group performance based on the features and evaluation of the prediction accuracy and model robustness.

Second, a comparative model is proposed for the evaluation of individual student performance in relation to the group performance. In collaborative learning, a grade is generally given not to each student but to each group and assessment of group learning is typically dominated by measures assigned after collaboration (Gress et al. 2010), where the performance of each group is normally measured by the quality of the solutions or products generated (Goggins et al. 2015). It is, however, quite useful for teachers to understand the hidden performance of each individual student within the group. Moreover, isolating the impact of collaboration styles from the individual student qualities on the expected group performance allows to quantitatively analyze the groupwork improvement over individual tasks attributed exclusively to the way the group collaborated. The comparative performance analysis of both individual students and groups not only confirms quantitatively the advantages of collaborative learning over individual study, but most importantly explains exactly the circumstances and conditions when specific patterns of collaboration are successful or unsuccessful and why.

Third, we intend to investigate the impact of group composition on learning outcome in collaborative learning. A key finding of this work is the observation that groups with mixed-gender and diverse skills and abilities tend to benefit more from collaborative learning compared to uniform-gender groups of students with similar skills. We claim these improvements can be explained by a combination of a deeper diversity of skills, knowledge, and abilities to generate creative content, as well as increased engagement and focus during group work, especially in cross-gender communication and interaction.

The major contributions of our work can be concluded as follows:

1) Identification and extraction of the factors and features from collaborative learning process that affect group performance;
2) Unified feature normalisation across diverse assignments and assessment methods;
3) Expressing diverse student's learning abilities through feature definitions;
4) Exploring group performance predictability based on live interaction dynamics in a form of application of classification and regression models (Extreme Learning Machine based Feed-Forward Neural Networks and Classification and Regression Trees) to group performance prediction;
5) Group performance prediction model validation on live data acquired from the trial carried out with 122 students;
6) Proposition of a new comparative model of individual student performance assessment in relation to and based on the group performance;
7) Quantitative definition of a group learning synergy expressed as a difference between the group actual assessment and its expectation that is made of the sum of individual members' contributions;
8) Investigating the impact of group composition on collaboration performance and providing quantitative measurable evidence of groups with mixed-gender and diversity of skills performing better as compared to uniform-gender groups of students with similar skill in collaborative learning.

It is important to state that this work and the above contributions focus on the prediction of group performance and other attributes of groupwork after completing the task. However, without any loss of generality, they can be applied at any stage during groupwork with the impact on predictive preformance proportional to the level and completedness of live tasks. In this respect, there are therefore no intrinsic limitations of applying the presented group performance prediction and knowledge discovery methodologies in real-time, even during live classroom activities.

It has been suggested in (Cress 2008) that analysis of CSCL should look into both the group effects and individual level and, from there, carry out multilevel analysis on the hierarchical structure of learning data. A Multilevel Model (MLM) has been proposed for CSCL, which allows different regression functions with different intercepts and different slopes for each of all groups in linear regression; as an example, the relation between satisfactions of individual students and their activity in collaborative learning is analyzed based on multilevel analysis (Cress 2008). Although it is quite efficient for CSCL analysis, it requires an enormous sample size, which may not be quite feasible in practice (Cress 2008). In our work, learning performance is assessed using both groups and individuals as the units of analysis by considering the hierarchical structure of group learning data in multilevel analysis. In group performance prediction, different learning abilities of individuals are expressed in feature representation and unified by feature normalization. The individual performance is assessed by comparing individual contributions to corresponding group workloads and the achievement of these groups in consecutive assignments with a deterministic comparative performance model. It has been demonstrated that our proposed methods can be applied in group learning analysis with a small CSCL data set. The results described in this paper seek to quantitatively prove the synergistic improvements of collaborative learning, evaluate their extent, and explain them in terms of the properties of student interaction and group diversity. We believe the findings arisen from this work can provide the education community with useful insights in organizing their own collaborative learning processes and student group structures in order to achieve optimal learning outcomes.

The remainder of the paper is organized as follows. Background on collaborative learning and related work on the three issues addressed in our work are introduced in Section 2. Section 3 introduces the CLE developed at EBTIC and discusses its features. Group performance prediction based on classification and regression models is presented in Section 4, where the diversity of assignments and students is considered in feature representation. A comparative analysis model to evaluate the performance of an individual student in a group and a generative mixture model of group performance are proposed in Sections 5 and 6, respectively. The quantitative results from the prediction experiments are shown in Section 7. Finally, the concluding remarks are given in Section 8.

Collaborative learning

Collaborative learning is defined by Johnson et al. as the instructional use of small groups so that students work together to maximize their own and each others learning (Johnson et al. 1991). In recent decades, various theories of how collaboration works for learning, which are associated with information exchange, conflict resolution, inter-subjective mearning-making, group knowledge building, and participatory models, have been proposed (Suthers 2006). In contrast to individual learning, collaborative learning is characterized as a field centrally concerned with meaning and practices of meaning-making in the context of joint activity, and the ways in which these practices are mediated through designed artifacts (Koschmann 2002). With the development of personal computers, mobile devices and wireless communication, CSCL, characterized by the sharing and construction of knowledge among participants using technology as their primary means of communication or as a common resource (Stahl et al. 2006), has been considered as an effective way to improve performance and efficiency of learning (Slavin 1990; Johnson et al. 2000). Significant changes in learning efficiency tend to be observed when students work collaboratively within groups rather than working individually, which is, in principle, attributed to being helped by partner students or helping partner students (Stahl et al. 2006). As an example, lower-ability students are reported to benefit much more from learning in a collaborative setting than higher-ability students (Saner et al. 1994). This observation matches the intuition that higher performing students on average tend to transfer knowledge to the lesser performing students.

Besides student academic performance, collaborative learning is also able to improve student interpersonal, intercultural and higher level thinking skills (Johnson and Johnson 1988; Slavin and Cooper 1999). Collaborative learning activities provide students with chances to explain their understanding of the subject matter to their group members, which can help students elaborate and reorganize their knowledge (Van Boxtel et al. 2000). It is also shown that discussion among students during collaborative learning can improve their ability of understanding and interpretations (Fall et al. 1997). Collaborative learning can also train students to work better in teams and to participate more effectively in a democratic society

(Feichtner and Davis 1991; Kagan 1994). It is even postulated that the experience achieved from collaborative learning is essential for the healthy psychological development of students (Johnson et al. 1998).

Unlike individual learning where the learning outcome of a student is dominated by his/her personal learning characteristics, e.g. learning ability, time spent, etc., effective collaborative learning involves not only the contribution of individual students but also depends on the way the group comes together to produce contributions. This may involve interdependence, concurrency, work distribution, mutual evaluation and reflection. It has been reported in the literature that the outcome of group-based learning in collaborative learning can be influenced by student characteristics, task characteristics, group composition, and team collaboration, e.g. positive interdependence, individual accountability, promotive interaction, social skills and group processing (Johnson and Johnson 1998; Lai 2011). To maximize the effectiveness of collaborative learning, the needs for students to be trained handling group issues (Oakley et al. 2004) and for teachers to be guided in training students on how to conduct group work (Ward 2006) have been highlighted. The importance of creating structured group assessments has been explored as well (Cohen et al. 2002; Vita 2005). The effectiveness of the student collaboration has a high impact on the learning outcome, which is dependent on the quality of interactions, especially the degree of interactivity and negotiability (Dillenbourg 2000). It has been shown that students' collaborative work on the same assignment followed many different interaction patterns, which can greatly affect the performance and assessment of the group work (Cen et al. 2014a). Continuous focus, self-reflection, live collaboration, and a fairly even distribution of workload and contributions are naturally more likely to lead to more refined and coherent assignment outcome, and consequently achieve better marks.

The behavior of a group is more than the sum of its individual parts, which indicates that group collaboration evolves in ways that are not necessarily evaluated based on the inputs of group members (Dillenbourg et al. 1996). This, in turn, brings much more complexity and challenges to the implementation of collaborative learning. Recent studies on collaborative learning have shifted the theoretical focus from individual functions within groups to an overall analysis based on whole groups (Dillenbourg et al. 1996). Although many studies on CSCL have been reported in the literature, more research is still required to achieve efficient learning implementation and practice. Quantitative analysis has played an important role in CSCL research to gain in-depth understanding of collaborative learning (Bruckman et al. 2002). In this work, several quantitative approaches have been developed to analyze the characteristics of collaborative learning and assess their impact on learning performance. Specifically, we focused on generic capability to estimate or predict group performance at different stages of the joint group learning exercise: before, during, and after the group task. Machine learning based approaches have been proposed to predict group learning performance during the exercise utilizing live members interactions and other dynamics describing concurrent and shared contributions. Effective and normalized features have been developed to provide the most explanative power against standardized actual assessment grades. Then, a simple prior group assessment expectation model that combines individual student performances has been updated by adding generative components that allow us to reliably predict group performances. The deviations between the predictive expectation model and the actual assessment provided by the teacher was coined as an objective quantitative measure of collaboration synergy that directly measures whether the collaboration is effective or not and whether it brings group performance gains or, in some cases, the opposite. Finally, effective group composition and its impact on group performance have been investigated, which reveals new quantitative insights as to the role gender distribution and skill-diversity in the group have on its performance. All analysis has been been carried out on the group coursework interaction data linked with the teacher-led assessments. All data collection was carried out using the EBTIC-developed CLE platform. The following sub-sections provide a thorough review of the related work reported in the literature.

Prediction of group learning performance

With the development of machine learning technologies, predicting students' academic performance, i.e. predicting how well a student will perform on a learning task based on historical knowledge or data (Thai-Nghe et al. 2011b), is one of the oldest and most useful applications of educational data mining (Romero et al. 2013), and has attracted increasing attention in the learning community. Student performance prediction could provide informed guidance, advice, and early feedback that may help to improve student's knowledge retention, formal assessment outcomes and satisfaction from the educational experience. A common observation that students enjoy dealing with and achieve successes in subjects they are naturally good at seems to support this claim. Furthermore, a good and reliable prediction model could, in the long-run, define student curriculum paths and possibly replace standardized examinations, thereby reducing exam pressure and workload which negatively affect both teachers and students (Thai-Nghe et al. 2011a; Feng et al. 2009; Thai-Nghe et al. 2011b).

This is, however, quite a challenging problem, since the learning performance of students can be cross-affected by lots of factors, e.g. demographic, cultural, social, or family factors, socio-economic status, psychological profile, previous schooling, prior academic performance, interactions between students and faculty, etc. (Romero et al. 2013; Araque et al. 2009). Many machine learning techniques, e.g. linear regressing (Feng et al. 2009), logistic regression (Cen et al. 2006), decision trees (Thai-Nghe et al. 2007; Yadav and Pal 2012), neural networks (Romero et al. 2008), support vector machines (Thai-Nghe et al. 2009), smooth support vector machine classification associated with kernel k-means clustering techniques (Sembiring et al. 2011), Bayes classification (Bhardwaj and Pal 2011), and matrix factorization based recommendation technique (Thai-Nghe et al. 2011a; Thai-Nghe et al. 2011b; Thai-Nghe et al. 2010), have been applied to solve student performance prediction problems in the literature, and, depending on the definition of the problems and the types of variables to be predicted, different techniques are employed, such as classification for categorical variables, regression for continuous variables or density estimation when the predicted values are probability density functions (Romero et al. 2013; Hämäläinen and Vinni 2011). An incremental ensemble of base classifiers, i.e. Naive Bayes, the 1-NN and the WINNOW algorithms using the voting methodology, has been proposed for predicting student performance in distance education (Kotsiantis et al. 2010). It is proposed for online learning to identify poor performance in an open and distance learning environment, where its data arrive continuously and it is impractical to store data for batch learning. A predictive analytic model has been developed for the University of Phoenix to identify students who are in danger of failing the course in which they are enrolled, based on timely intervention strategies (Barber and Sharkey 2012). Three models have been developed to predict student failure for distance learning by analyzing the clicking behavior of students in a virtual learning environment (Wolff et al. 2013). A prediction of the students' grades for assignments they are currently undertaking is made by monitoring students' progress based on their participation in online collaborative learning activities. Selective interventions are then taken to prevent the students from actually failing

(Gunnarsson and Alterman 2012). Students' final performance is predicted by using different data mining approaches based on participation in online discussion forums that constitute communities of people learning from each other (Romero et al. 2013). Instead of using traditional classification, it applies clustering plus class association rules mining to build student performance models. It has been illustrated from the results of experiments conducted for the first-year computer-science university students, that the approach is suitable for performing both a final prediction at the end of the course and an early prediction before the end of the course.

In collaborative learning, the learning behavior of students working collaboratively is more complicated than that of individual learning (Hackman and Morris 1975). The performance of a group is not decided by individual learners, but is a complex combination of all learners' contributions to the group. Assessment and prediction of group performance can help to evaluate and improve a collaborative learning system, identify productive grouping and interaction patterns, and help to understand what drives student academic performance within a dynamic and connected learning environment. As mentioned before, both the characteristics of individual students and their interaction patterns can influence the performance of group learning, which makes performance assessment and prediction in collaborative learning much more challenging compared to individual learning.

After reviewing 186 papers and 340 measures, Gress et al. classified group performance assessment in collaborative learning into categories of self-report, interview, observation, process data, discussions and dialogues, performance and products, and feedback (Gress et al. 2010). However, these assessment methods, although commonly used, tend to violate the assessment requirements of time, validity, reliability, and individual accountability in one way or another (Xing et al. 2014). Examination of final products of collaboration, e.g. group assessment, by using the average score of outcome of each task (Zhu 2012), or the quality of the solution produced according to a holistic rubric with consideration of productive failure in collaborative learning for ill-structured and well-structured problems (Kapur and Kinzer 2009), has been a dominant means to evaluate group learning performance (Goggins et al. 2015). However, an assessment based solely on learning outcomes cannot accurately measure group performance since it overlooks elements of the collaborative learning process (McNely et al. 2012; Strijbos 2011), such as group dynamics, interaction, and technology-mediated processes (Goggins et al. 2015). To address these aspects, qualitative methods have been developed for assessing group performance based on team collaboration indicated from its dialogue during collaboration (Safin et al. 2010). The major disadvantages of these methods are that they are time-consuming and difficult to implement. Some quantitative methods have been proposed to overcome these drawbacks by quantifying complex collaborative processes either by building ad-hoc measures or by quantifying categories of actions or utterances (Goggins et al. 2015; Strijbos 2011). For example, quantitative content analysis has been used to characterize group discussion by coding and counting the frequencies of different aspects of discourse (Kapur et al. 2011). However, these approaches cannot accurately define the learning process and group collaboration in a quantitative way (Goggins et al. 2015). The study by Xing et al. (2014) assesses CSCL by using activity theory, where an automated strategy is proposed to assess participation in a multi-mode math discourse environment called Virtual Math Teams with Geogrebra (VMTwG). Most studies that are based on statistical modeling and data mining techniques are focused on methodology, exploration of algorithms and mathematical modeling, in ways tending to overlook educational contexts, theories and phenomena (Xing et al. 2015; Baker and Yacef 2009; Romero and Ventura 2010). To overcome this limitation, a methodology which connects perspectives from learning analytics, educational data mining, theory and application to predict students' performance in the VMTwG environment with small datasets has been proposed (Xing et al. 2015). In this method, students' participation in a CSCL course is holistically quantified based on activity theory, and learning performance is predicted by using Genetic Programming (GP) based prediction model. In the literature, other approaches for collaborative learning analytics based on different machine learning technologies have been proposed as well, e.g. Bayesian networks, decision trees, and fuzzy logic (Ferguson and Shum 2011; Coffrin et al. 2014). A regression prediction strategy is proposed to predict group performance according to the students' functional roles which are identified automatically based on the analysis of their online collaborative learning interactions (Coffrin, Corrin, Barba, & Kennedy 2011). A prediction model is formalized by using system-tracked data to forecast team performance, where the log records are analyzed to measure group and individual participation, and direct and indirect measures of involvement are used as predictor variables (Goode and Caicedo 2014).

In this work, novel prediction models based on supervised learning techniques are proposed for group performance prediction in CSCL using historical and live group interaction evidence. Compared to the techniques in the literature, the major advantages of our methods include: 1) definition of a number of discriminative features from concurrent sequences of student learning and shared content creation sessions to measure various characteristics of their contributions to joint assignments and their interaction within groups, and to analyze their abilities to differentiate between the likely outcome represented by the formal group assessment; 2) address of the challenges posed by accommodating different students, diverse assignments and assessment methods which are resolved through normalised and unified assessment representation and generic feature definitions; 3) application of both classification and regression models to satisfy different prediction goals.

Individual assessment

In collaborative learning, an assessment grade is generally assigned to each group based on the group's achievement, i.e. the quality of the solutions or products generated after collaboration (Goggins et al. 2015; Gress et al. 2010), which is then, in turn, assigned to all individual group members. It is useful as well to understand each student's individual performance in creating the final assignment. Learning within groups makes it difficult to isolate individual contributions and to assess the learning outcome of an individual from the group achievement. The final grade given to a group for an assignment created collaboratively does not necessarily reflect any one individual's effort, knowledge, skill, or ability, since students in the group may not make comparable or equivalent contributions (Saner et al. 1994; Race 2001; Webb 1995). Collaboration among students within groups can have evident effect on learning performance even with limited interaction, e.g. a 10-minute discussion (Fall et al. 1997). It has been demonstrated that group assessments may not accurately reflect individual achievements in collaborative learning (Saner et al. 1994; Webb 1993; Webb et al. 1998). Especially in heterogeneous groups with students having various ability levels, low-ability students may obtain higher grades from group assessments that are achieved based on the contributions of their high-ability teammates (Saner et al. 1994). If the higher performance of a group can reflect actual learning progress, the group assessment is not necessarily invalid; while if low-ability students are assigned higher scores based on the group achievement completed mostly by the higher-ability students in the group, assessing individual student learning using group performance as the indicator will not be accurate enough. Webb (1995) has suggested using individual student assessment instead of group-based assessment if the assessment of individual student performance is more important than that of the group. Although this provides accurate assessment of the individual students, the effectiveness and the synergy of groupwork will remain unassessed as a result, or, in extreme cases, the group may miss out on the benefits of collaborative learning altogether. In addition, it has been demonstrated that the achievement of group members are not independent of one another due to cooperation impact, and such impact can continuously affect the performance of subsequent individual work, especially for lower-ability students (Saner et al. 1994; Webb et al. 1998).

Most of the work in learning assessment is focused on student performance assessment in individual learning and group performance evaluation in collaborative learning. Individual student performance assessment in collaborative learning, although being an important measure for evaluating learning achievement, has not been widely addressed in the literature. A negotiation model has been used to represent student interactions for assessing the performance of individual students in collaborative learning (Dillenbourg et al. 1996). The interaction of students within groups is indicated by conversation strategies for students' learning assessment (Roschelle and Teasley 1995). Learning behaviors are analysed for assessing individual student interaction and performance (Webb 1991; Webb 1993; Webb et al. 1998). However, neither the interactions nor the student behaviors were quantitatively defined and analyzed in these methods. To address this, a comparative model is proposed for the definition and quantitative evaluation of individual student performance in relation to the group performance by considering the characteristics of learning and the relationship between contribution and achievement.

Group composition

A learning group is usually characterized by its size, gender and ability levels of its individual members assuming all members have similar ages. The size of a group indicates the amount of knowledge exchange and collaboration available during the learning and content generation process. In general, larger groups with reasonable sizes tend to perform better than smaller ones if learning activity levels and individual characteristics are similar (Cen et al. 2014b). It has been noted in the literature that Q18 the composition of groups with different abilities and genders of students is closely related to the ways students engage, collaborate and learn (Webb et al. 1998; Webb 1991; Savicki, Kelley, & Lingenfelter 1991; Savicki et al. 1996; Gordon 2000), which Q19 consequently influences learning performance. However, few quantitative approaches are provided in the literature for analyzing group composition.

Webb (1991) has shown that both interaction patterns and collaboration effects can Q20 vary across groups with varying ability-level compositions. Specifically, high-ability students in groups tend to contribute more by providing more explanations and information while low-ability students are more likely to be off-task (Webb 1991). Low-ability students having high-ability peers as teammates are more likely to significantly improve their performance on both group tasks and individually-completed post-tests, while the performance of high-ability students working in heterogeneous groups consisting of students with varying ability-levels is not affected by the group composition (Webb et al. 1998). Although there is no quantitative measurement of performance improvement and ability-levels, this finding provides us with a strong recommendation for encouraging groups' heterogeneity in collaborative learning. It has also been observed (Webb 1991) that in heterogenous groups with wider ability range, higher-ability and lower-ability students tend to form teacher-student relationships and interact and collaborate more among themselves, and as a result medium-ability students tend to be left out.

Significant research has been dedicated to study and compare the effectiveness of single-gender education and mixed-gender education, i.e. co-education with various gender compositions in classrooms. The focus of these studies aimed to employ gender-specific educational strategies for varying purposes such as enhancing student confidence and skills, improving learning outcomes, or towards achieving social mobility (Zeid and El-Bahey 2011). The difference in the results of these studies is mainly caused by gender specific characteristic and behaviors in collaborative learning. The study carried out on social mobility (Zeid and El-Bahey 2011) indicates that females tend to focus more on socially oriented activities while males tend to focus more on task-oriented activities. Moreover, female students learning together in a technology-rich environment seem to participate more actively and persistently than male students regardless of the nature of the task (Goldstein and Puntambekar 2004). Similarly it has been found that female engineering students collaborate more often as a successful learning strategy compared to their male classmates (Stump et al. 2011). The research outcomes (Chennabathni and Rgskind 1998) claim that girls in high schools perform better with single-gender groups when learning unfamiliar tasks but excell more in mixed gender groups when learning familiar tasks. The study (Webb 1991) conducted with groups of mixed genders shows that girls are more likely to be ignored by their boy teammates and fail to acquire answers to their questions when majority of the members in a group are male. Zeid and El-Bahey (2011) found that the overall course performance for both genders was improved by changing the software engineering classroom composition from a gender heterogeneous to a gender homogeneous classroom. However, it has been found that mixing different genders in one learning group could possibly arouse learning enthusiasm of students who are willing to contribute more in the learning process (Cen et al. 2014a). Although the effects of single-gender education and co-education have still been disputed with contradictory opinions, e.g. (Mael et al. 2005; Morse 1998; Crosswell and Hunter 2012; Smith 1996), these studies indicate that gender composition can largely influence learning outcome in collaborative learning.

For quantitative analysis of group composition, a generative mixture model is proposed in our work to isolate the impact of collaboration style from individual student qualities on group performance. Group composition across genders and diversity of skills and abilities is quantitatively analyzed based on this mathematical model.

Collaborative learning environment

The Collaborative Learning Environment (CLE) is a system developed at EBTIC (Ruta et al. 2013; Hirsch et al. 2013) that brings together a collection of tools and functionalities enabling communication, information sharing and collaborative document creation within the same environment. As opposed to individual communication and sharing tools like Skype, Facebook, or Google Drive which focus on a specific interaction or activity, the CLE is designed to integrate these different functionalities together into one cohesive learning environment. The CLE has been used to collect most of the data that are used in the remainder of the article.

The CLE is implemented as a set of modules for Moodle, an open source learning management system (LMS), and as such is able to capitalise on existing Moodle functionalities like group creation, file sharing and forums. By leveraging the flexibility of open-source technologies, the CLE integrates seamlessly into the LMS, providing a workspace that is familiar to both students and faculty, thereby reducing cognitive load and enabling more focus to be placed on the collaborative process.

The aim of the CLE is to stimulate the collaborative learning process and enable instructors to facilitate collaborative assignments more easily. Moreover, the whole interaction history is logged, which provides data enabling dynamic analysis of contributions, usage and participation, as well as enabling more advanced future functions such as knowledge elicitation. A screenshot of the CLE is shown in Fig. 1.

Communication features of the CLE include synchronous text chat and audio/video communication, which allow participants to exchange ideas and communicate directly with each other regardless of their geographic location. Additionally, a collaboration area is provided to allow students to either synchronously or a-synchronously create an assignment. This area, called the collaborative editing pad, provides a canvas on which students can contribute and revise their ideas. Each contributor to the pad is assigned a unique color, so individual contributions are evident, and each keystroke, whether it is an "add", "edit" or "delete", is recorded by the pad.

The writing area of CLE is powered by etherpad-lite, a real-time collaborative text editor. Students edits are collected and stored about 60 times per minute. Etherpad-lite stores the change-sets in its database, associated with a timestamp, user, and pad-id. CLE then extracts relevant details (author, assignment, group, time of change, change-type (addition, deletion, copy/paste)) for further analysis.

All individual students' assignment progress time series are merged into a single colour-coded progress timeline that can be viewed and played back like a movie. The students'

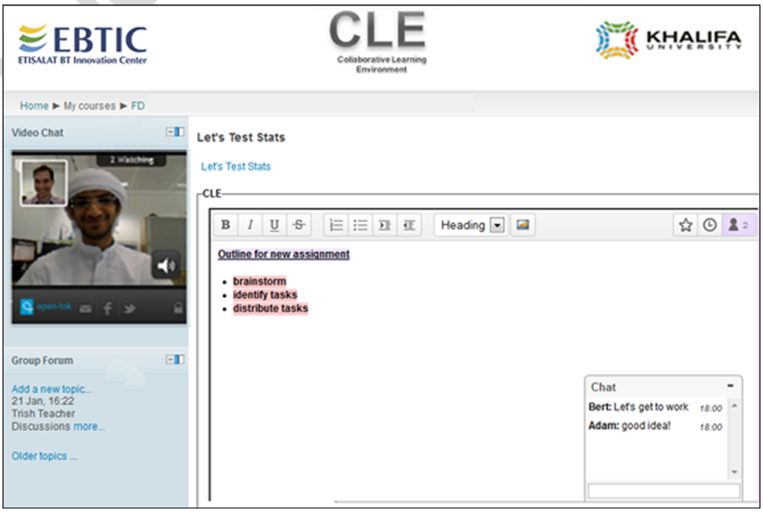


Fig. 1 A screenshot of Collaborative Learning Environment (CLE) in action

collaborative work on the same assignment follows many different patterns, from sequential to concurrent contributions, from one person dominated to evenly distributed workloads, from continuous progression of contributions to sudden bursts of activity and/or paste-ins. Figure 2 illustrates several examples of progress timelines, signifying different patterns of group interactions while working on a single assignment. Such progress timelines are constructed by measuring the cumulative volume of keystrokes by different students, marked in different colors, along the time while working on the group coursework. The vertical axis measures the cumulative volume of the assignment content changes recorded by the CLE, and the horizontal axis represents the time stamps of these changes up to 1 *s* resolution. These progress timelines provide a summarised view of how the contributions from each group member evolved over time. It has been revealed in our prior study, exploring the impact of students collaborative work on students performance, that continuous focus, self-reflection, live collaboration, and fairly even distribution of workload and contribution are more likely to lead to more refined and coherent assignments, and consequently achieve better marks. These findings are important to identify discriminative features in learning performance assessment and prediction.

Another way of illustrating students' interaction data is by depicting the volume of edit activity (measured in keystrokes) at a higher time resolution level. An example of such an edit activity plot is depicted in Fig. 3. These two illustrations are just some examples of the interesting insights into how the group collaborated together. There are many other aspects of collaboration within CLE that can be analysed, for example monitoring the exact locations of

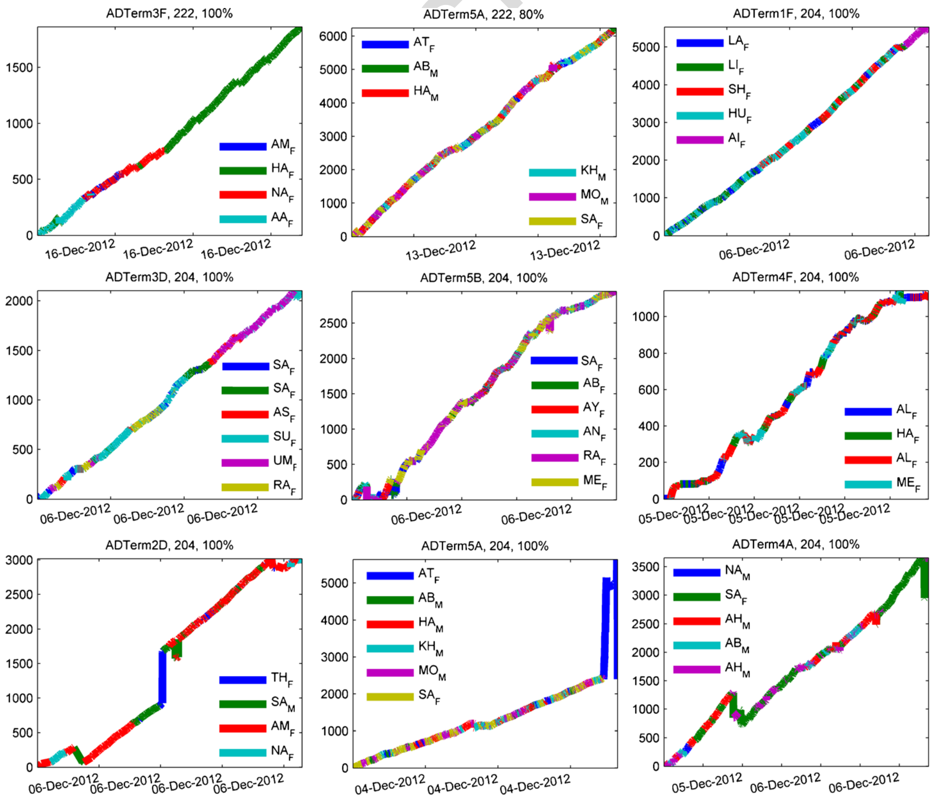
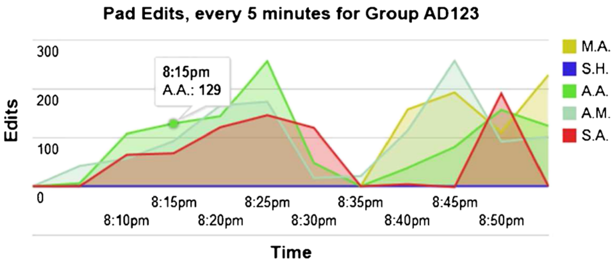


Fig. 2 Sample patterns of student interactions while working on group assignments: progress timelines

Fig. 3 CLE statistics module, sample group contributions graphs



different contribution edits to assess whether the members are working independently on the same or different parts of the assignment or actually trying to understand and assess the value of individual contributions.

All the above approaches to analyse collaborative learning data harvested by the CLE are supported by a statistics module that can output detailed usage statistics at different aggregate levels. Such insights are invaluable for the instructors as they allow them to obtain, in an instant, a detailed analysis of how the group assignment was completed and what were the individual members' effort and actual contributions. Beyond statistics, the CLE also provides a playback feature, that allows both the students and instructors to watch the entire creation of the assignment, from start to finish, much like watching a video.

Group performance prediction

This section describes group performance prediction in CSCL, in which the extraction and normalization of features representing contribution and interaction of students and the machine learning based prediction models are proposed. Machine learning, as a type of artificial intelligence (AI), aims at developing algorithms that provide computers with ability to learn from historical data and make data-driven predictions or decisions without following strict and explicit program instructions. Depending on the nature of the data, machine learning tasks are typically classified into three broad categories, i.e. supervised learning, unsupervised learning, and reinforcement learning. Supervised learning aims at inferring a function from a set of training examples, each of which consists of an input object (features typically represented by a vector) and a desired output (label), while unsupervised learning aims at discovering hidden patterns in training data without labels provided to learning algorithms. In reinforcement learning, a computer program interacts with a dynamic environment in which it performs a certain goal, e.g. driving a car or playing games, without explicit instruction on whether it has come close to its goal or not. According to the values of the target variable, supervised learning has two categories; classification where target variables are categorical, and regression where target variables are continuous. In our work, group performance prediction is formulated as a supervised learning process and both classification and regression models are built for the prediction task. In classification and regression, a feature is an individual measurable property of a phenomenon being observed (Bishop 2006). Successfully solving classification and regression problems is largely dependent on the choice of informative and discriminating features. Features are, in general, numeric, as used in this paper, while structural features such as strings and graphs are also used in syntactic pattern recognition. Successful approaches are developed for feature extraction and normalization to capture statistics of the contribution and interaction of student members within groups during the learning process. These features are used to learn the classification labels or regression outputs defined as normalized grades awarded to the groups as their formal assignment assessment. This will be elaborated further below.

Classification and regression models

Classification and regression are supervised learning techniques to build prediction models from data. They share a common framework, which is shown in Fig. 4. During the training process, a feature set is extracted from the training data to capture important and discriminative attributes of input data in relation to the target attributes. Pairs of feature sets and given target values are then fed to the learning algorithm to construct a mapping between the data features and the target attributes, which constitutes model learning. Once the model is built, the same feature extraction is applied to unseen data and the learned model uses these data to predict the target variable.

The main difference between classification and regression is the representation of the target variable. In classification, the target variable is categorical, taking values associated with different possible classes of output. Whereas, a regression model has a numerical and, usually, continuous output variable. In this section, the academic performance prediction is first formulated and described as a classification problem, and then a regression problem.

When learning performance prediction is formulated as a classification problem, the outputs of the target grade variable are discretised into just several possible grade levels taking integer values that are set to be within [1,5] in our method. These levels correspond to grades and are used as class labels during model learning and prediction. Note that in such defined classification the prediction error is also granular, i.e. it is either 0 or equal to the difference between the actual and wrongly predicted grades.

In statistics, regression analysis is a methodology for estimating the relationships between a dependent variable and one or more independent variables, which has been widely applied in prediction and forecasting. When building a regression model for performance prediction, the dependent variable represents group grades and the independent variables are the data features. Unlike classification, the grades are normalized to take continuous values within the range of [1,100]. The advantage of using a regression model is that it provides a finer granularity of the target variable and normally better reflects similarity between different target values simply by their distance.

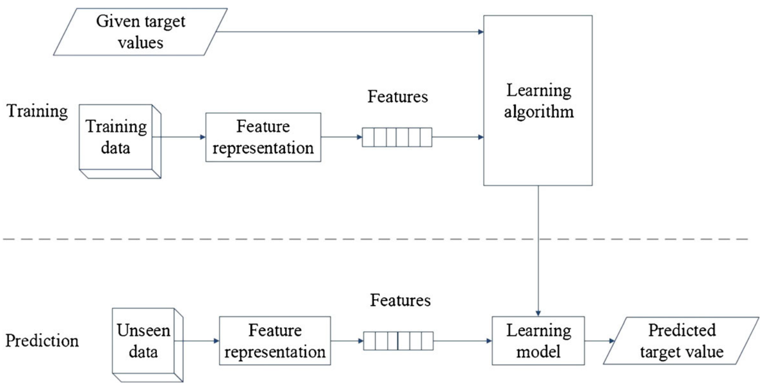


Fig. 4 Framework in supervised classification and regression

As shown in Fig. 4, feature representation and model learning are two major parts in both classification and regression frameworks. Typically they consume most of the time spent in building the classification or regression model.

Feature representation

The features used for grade prediction are directly extracted from students' interaction within groups recorded along the duration of the assignments. The data collected for an individual student during each assignment are listed below:

1. ID and first name of the student;
2. ID and name of the group that the student belongs to;
3. ID of the assignment that the student is completing;
4. Indices of revisions reflecting the order in which they were updated;
5. The contribution attributes: start length, end length, the total number of changes and the volume of length changes, as well as the time stamp of the contribution;
6. Grade awarded by the teacher for the assignment which, in this case, is the assessment for the whole group.

The following list captures some simple aggregate statistics derived from the above data:

1. ID and name of the group;
2. The number of members in the group;
3. The number of revisions carried out by each group member;
4. The total number of revisions carried out by the whole group;
5. The absolute number of changes (adding/deleting), the number of positive changes (adding), and the number of negative changes (deleting), carried out by each group member;
6. The absolute number of changes (adding/deleting), the number of positive changes (adding), and the number of negative changes (deleting), carried out by the whole group;
7. Group performance, i.e. group grade.

It can be seen from the recorded data that the number of revisions and the total number of changes in all revisions represent the effort made in learning, which have been observed to be closely associated with learning outcome (Cen et al. 2014b). Let $ratioRL_{ga}$ be the number of revisions in the unit length of the change made by the *gth* group in the *ath* assignment, which is calculated as

$$ratioRL_{ga} = R_{ga} / \left(L_{ga1} - L_{ga2} \right), g \in G, a \in A, \tag{1}$$

where $_G$ and A denote the sets of indices of groups and assignments respectively, R_{ga} is the number of revisions, and L_{ga1} and L_{ga2} are the start length and end length of the total changes made by the *gth* group in the *ath* assignment. Considering that there are large differences among efforts made in various revisions, we use the ratio of $ratioRL_{ga}$ instead of using the number of revisions as features. Q22

As described before, there are positive changes and negative changes, reflecting the adding and deleting of content respectively. Positive changes represent valid contributions that are used in the subsequent work, while negative changes indicate that some of previous contribution are considered to be invalid. Since content editing can be carried out synchronously or asynchronously by different student members within a group, the amount of both positive and negative changes indicates the interaction among students and provides a way to estimate learning efficiency. The absolute number of changes measures effective changes remained in the final evaluation, which indicates group contributions. This together with the numbers of positive changes and negative changes is then used as features in our method:

$$C_{ga} = \left\{C_{|ga|}, C_{+ga}, C_{-ga}\right\}, g \in G, a \in A, \tag{2}$$

where C_{ga} denotes the change vector of the gth group in the ath assignment, $C_{|ga|}$ is the absolute number of changes, and C_{+ga} and C_{-ga} are the positive and negative changes respectively.It has been shown in our previous work (Cen et al. 2014a) that the size of a group is also an important factor affecting the group learning outcome. It indicates the amount of knowledge exchange and collaboration available during the learning and content generation process. Let NS_{ga} denote the number of students in the gth group taking the ath assignment. The features used in prediction can then be expressed as

$$f_{ga} = \left\{NS_{ga}, ratioRL_{ga}, C_{ga}\right\}, g \in G, a \in A. \tag{3}$$

In (3), f_{ga} represents the feature vector of the gth group in the ath assignment, which quantifies collaboration and interaction made by this group in the learning process of a assignment.

However in practical applications, there are often multiple assignments in one course, as is the case in our work in which there are 3 assignments. Since the contents and tasks in different assignments can be totally different, the features representing contribution and interaction given in (3) are normalized by considering the diversity of assignments. To implement this, a weighting coefficient is allocated to each of assignments to measure its difficulty level. Let $\overline{grade_a}$ and $\overline{ctr_a}$ be the average grade and average contributions made in the ath assignment, respectively, which can be expressed as

$$\overline{grade_a} = \frac{\sum_{g=1}^{NG_a} grade_{ga}}{NG_a}, a \in A, \tag{4}$$

and

$$\overline{ctr_a} = \frac{\frac{\sum_{g=1}^{NG_a} C_{|ga|}}{NG_a}}{\frac{\sum_{g=1}^{NG_a} R_{|ga|}}{NG_a}} = \frac{\sum_{g=1}^{NG_a} C_{|ga|}}{\sum_{g=1}^{NG_a} R_{|ga|}}, a \in A, \tag{5}$$

where $grade_{ga}$ is the grade achieved by the gth group in the ath assignment, and NG_a is the number of groups taking this assignment. The difficulty weight of the ath assignment, denoted as $diff_a$, is then calculated as

$$diff_a = \frac{\overline{ctr_a}}{\overline{grade_a}}, a \in A. \tag{6}$$

It can be seen from (6), the assignment with higher value of $diff_a$, is likely to have lower average grade and require more contributions, and is thus considered to be more difficult. It should be noted that this can apply only if all the assignments are involved in one course and taken by the same students, in which all comparisons can be made with a unique reference. Here, $diff_a$ is normalized within [0,1]. The revisions and changes in each assignment are then normalized according to $diff_a$, and the features in (3) are as such re-written as

$$f_{ga} = \left\{NS_{ga}, ratioRL_{ga} \times diff_a, C_{ga} \times diff_a\right\}, g \in G, a \in A. \tag{7}$$

By doing so, the group contributions are normalized based on the assignment difficulty, hence allowing different assignments to be compared at a similar level.

Although the assessment grade is allocated to a whole group, the efficiency, knowledge background and level of understanding of the participating students are different. To differentiate among individual students' performance in the feature representation we allocate a performance weight to each individual student in the group. Let w_s be the performance weight of the sth student, which is calculated based on the average grade achieved in all assignments completed by this student and is given as

$$w_s = \frac{\sum_{a=1}^{NA_s} grade_{ga}}{NA_s}, s \in S, \tag{8}$$

where S is the set of indices of students, NA_s is the number of assignments that the sth student has completed, and $g \in G$ is the index of the group that the sth student belongs to in the ath assignment.We then estimate a possible achievement of a group by considering the contributions and performance weights of all member students in the group, shown as:

$$\hat{P}_{ga} = \sum_{s=1}^{NS_{ga}} C_{|sa|} \times w_s, \quad g \in G, \quad a \in A, \tag{9}$$

where $\hat{P}_{ga}$ is the performance expectation of the gth group in the ath assignment, and $C_{|sa|}$ is the absolute number of the changes made by the sth student in the ath assignment. The $\hat{P}_{ga}$ is normalized within [0,100] to satisfy the purpose of having a finer granularity, which is added as a feature in (7). Since the calculation of $\hat{P}_{ga}$ has taken into account the number of changes made during assignments, it will not be used again. The feature, f_{ga} in (7) can therefore be changed to:

$$f_{ga} = \left\{NS_{ga}, ratioRL_{ga} \times diff_a, \hat{P}_{ga} \times diff_a\right\}. \tag{10}$$

In this way, the features are further normalized to eliminate the effect of different tasks and various learning abilities of different students. It is important to reiterate at this point that all the feature definitions and the related group performance prediction analysis, described above, are applied to completed groupwork task, yet it can also be equally applied in the exact form for any moment live during the groupwork. In this case the features would be recalculated continuously on a cumulative basis during the progress of the groupwork, and the predictive performance might be affected in proportion to the completeness of the groupwork interaction and its consistency of impacts on performance.

Learning algorithms

The problem of academic group performance prediction, formulated in this work, is solved using classification and regression models as selected instances of machine learning techniques. Specifically, neural networks and decision trees have been chosen for this task.

Extreme learning machine (ELM) based feedforward neural networks (NN)

Traditional feedforward neural networks extensively use slow gradient-based learning algorithms to train neural networks and tune the parameters iteratively, which makes their learning speeds rather slow. To overcome these drawbacks, Huang and his colleagues have proposed a new learning algorithm called extreme learning machine, which randomly chooses hidden nodes and analytically determines the output weights of the network (Huang et al. 2006). Compared to other computational intelligence methods such as the conventional back-propagation (BP) algorithm and support vector machines (SVMs), the ELM has much faster learning speeds, ease of implementation, least human intervention, and high generalization performance. It has been reported by Huang et al. (2006) that the ELM can produce better generalization performance and can learn thousands of times faster than traditional learning algorithms for feedforward neural networks.

Classification and regression trees (CART)

CART introduced by Breiman and his colleagues (Breiman et al. 1984) have been widely used in data mining and machine learning. It is used to build a model that is able to predict the value of a target based on the values of input attributes. The prediction models are constructed from data propagating through the condition tree until the leaf is reached. Specifically, the models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. Binary trees are constructed by repeatedly splitting a node into two child nodes, beginning with a root node that contains the whole learning sample, which are used for predicting categorical target variables in classification or continuous output variables in regression.

Comparative student performance model

Generally in collaborative learning, one assessment is allocated to a whole group, which is measured by the quality of the solutions or products generated after collaboration (Gress et al. 2010; Goggins et al. 2015). However, it is quite useful to understand the performance of each individual student. This is not only helpful to assess the academic achievement gained by each student, but also helpful to analyze the impact of collaborative learning. To address this, a comparative student performance model is proposed in this section.

In our method, the performance of an individual student within a group achieved for a particular assignment is modeled based on the grade of the group this student belongs to in this assignment, the other assignments he/she has completed before, and his/her contributions made to the corresponding assignments. Specifically, if the group marks are higher while the contributions of the student are lower (or vice versa), the student is considered to have a lower individual performance than his/her group performance. On the other hand, if the group marks and the contributions of the student are both in-line for different assignments, the student is likely to be among top students in the group and assumed to have a higher performance than that of his/her group. This will be elaborated below.

Let g_1 and g_2 be the group indexes of the $\tilde{s}^{th}$ student in the a_1^{th} and a_2^{th} assignments respectively, and assume that the a_1^{th} assignment is completed before the a_2^{th} assignment. It is not necessary that the two groups contain the same student members. Now we will model the performance of the student in the a_2^{th} assignment by estimating the grade that the student may achieve if he/she individually completes the assignment without collaboration with the the other students.

Based on the relationships between $grade_{g_1 a_1}$ and $grade_{g_2 a_2}$, and between $C_{|\tilde{s}a_1|}$ and $C_{|\tilde{s}a_2|}$, there can be four different cases:

$$grade_{g_1 a_1} \le grade_{g_2 a_2}, \quad C_{|\tilde{s}a_1|} \ge C_{|\tilde{s}a_2|} \tag{11}$$

$$grade_{g_1 a_1} \ge grade_{g_2 a_2}, \quad C_{|\tilde{s}a_1|} \le C_{|\tilde{s}a_2|} \tag{12}$$

$$grade_{g_1 a_1} < grade_{g_2 a_2}, \quad C_{|\tilde{s}a_1|} < C_{|\tilde{s}a_2|} \tag{13}$$

$$grade_{g_1 a_1} > grade_{g_2 a_2}, \quad C_{|\tilde{s}a_1|} > C_{|\tilde{s}a_2|} \tag{14}$$

In the cases given in (11) and (12), the student is likely to have more negative effect on group performance, while the cases in (13) and (14) indicate that the contributions of the student are important to the group performance and his/her achievement tends to be better than that of the group.

The value of $\hat{grade}_{\tilde{s}a_2}$ is estimated based on $grade_{g_2 a_2}$ as

$$\hat{grade}_{\tilde{s}a_2} = grade_{g_2 a_2} + \Delta grade_{\tilde{s}a_2}, \tag{15}$$

where $\Delta grade_{\tilde{s}a_2}$ is the grade adjustment and calculated as

$$\Delta grade_{\tilde{s}a_2} = \left(grade_{g_1 a_1} - grade_{g_2 a_2}\right) \times \left(ctr_{\tilde{s}a_1} - ctr_{\tilde{s}a_2}\right). \tag{16}$$

In (16), $ctr_{\tilde{s}a_i}$ with $i \in [1,2]$ denotes the percentage of the contribution of the $\tilde{s}^{th}$ student to the a_i^{th} assignment, which is calculated based on the ratio of the absolute number of changes as:

$$ctr_{\tilde{s}a_i} = \frac{C_{|\tilde{s}a_i|}}{C_{|ga_i|}} = \frac{C_{|\tilde{s}a_i|}}{\sum_{s=1}^{NS_{ga_i}} C_{|sa_i|}}. \tag{17}$$

If more than two assignments were completed by the student, i.e. $a_2 > 2$ and $a_1 \in [1,2,\ldots,a_2 - 1]$, the assignment to which the student made the most contributions is chosen as a_1 in the model.

For the first assignment in a multi-assignment course or the sole assignment in a single-assignment course, the performance of an individual student is modelled based on the percentage of his/her contribution within the group. The basic idea is that the performance of a student is expected to be higher than the group performance if the contribution of the student is greater than the average student contribution in the group, and vice versa. Let $\overline{ctr_{ga}}$ be the average student contribution of the gth group in the ath $(a=1)$ assignment, which can be calculated as

$$\overline{ctr_{ga}} = \frac{\sum_{s=1}^{NS_{ga}} C_{|sa|}}{NS_{ga}}. \tag{18}$$

Let $\Delta ctr_{\tilde{s}ga}$ be the normalized deviation between $\overline{ctr_{ga}}$ and the individual contribution made by the $\tilde{s}^{th}$ student, which is expressed as

$$\Delta ctr_{\tilde{s}ga} = \frac{C_{|\tilde{s}a|} - \overline{ctr_{ga}}}{\sum_{s=1}^{NS_{ga}} C_{|sa|}}. \tag{19}$$

The grade adjustment is then defined as

$$\Delta grade_{\tilde{s}a} = \Delta ctr_{\tilde{s}ga} \times grade_{full}, \tag{20}$$

where $grade_{full}$ denotes the full mark, i.e. 100. The expectation of the student grade is then calculated with the sum of $grade_{ga}$ and $\Delta grade_{\tilde{s}a}$.In order to constrain students' expected grades within a reasonable range, an upper and lower bounds are set as:

$$grade_{ga} - 10 \le \hat{grade}_{\tilde{s}a_2} \le grade_{ga} + 10. \tag{21}$$

If $\hat{grade}_{\tilde{s}a_2}$ exceeds given range, it will be set to the closest bound.

Group composition

As mentioned before, group composition can considerably influence the learning outcome in collaborative learning, since different configurations of groups yield different collaboration patterns and learning behaviors. In this part, the impact of collaborative learning on groups with homogeneous- or heterogeneous-genders and abilities is analyzed based on the comparative student performance model given in Section 5. This is quite important since it allows for comparisons of learning outcomes among different group composition configurations.

To isolate the impact of collaboration style from individual student qualities on group performance, a generative mixture model is proposed. The model assumes that the grade of the assignment of the group is generated as a combination of the students grade expectations, improved or degraded by the collaboration type that the students choose to follow. Specifically, it is expressed as a linear combination of contribution-weighted performances of all individual students in the group:

$$\hat{grade}_{ga} = \sum_{s=1}^{NSga} \hat{grade}_{sa} \times ctr_{sa}, \tag{22}$$

where $\hat{grade}_{ga}$ is the estimated grade of the gth group in the ath assignment and $\hat{grade}_{sa}$ denotes the individual performance of the sth student in this assignment. Here, $\hat{grade}_{sa}$ is estimated by using the comparative student performance model described in Section 5. The deviation between the group performance expectation and the actual grade received is likely linked to the way the students collaborated together in the group. The impact of collaborative learning on various group composition configurations can then be quantitatively analyzed and compared. The analysis of gender and ability composition within the CLE platform throughout the trial will be presented in Section 7.

Experiment results

Description of experiment data

The data used in the experiments were collected via the CLE platform trialed during the Fall Semester of 2013. During this trial, CLE was used in two courses, the Molecular Biology Engineering Course and the Freshman Design Engineering Course. The CLE trial consisted of 3 collaborative writing assignments related to the students' end-of-term project. The end-of-term project began by splitting students into teams with each team choosing a project to complete based on a set of project proposals submitted by the faculty from varying disciplines around the university. The faculty then became the 'client' or 'customer' for the students to build and solve the proposed problem. The CLE assignments consisted of three collaboratively written parts at various stages of the project process. The first assignment had each team create a 'Team Charter' which would outline the structure of the team and the expected behavior for the team and each of its members. The second assignment, the 'Revised Client Statement', required each team to write an analysis of their client's problem based on the information they had gathered from the meetings with their 'client'. As stated in the assignment, students needed to reflect their new understanding of the design problem as a result of working through the conceptual design cycle. The third assignment asked the students to create their 'Final

Design Report' document using the CLE. It should be noted that, due to limitations in the CLE such as inability to add images, using the CLE for this assignment was optional. The assignments were all collaborative writing assignments, but collaboration was not closely monitored during this period by the faculty as the trial was focused on being a data gathering exercise.

In total, 168 students used the tool. For the purposes of this article, a subset of the data was used. First, only the Freshman Design Engineering Course was taken into consideration. While 152 students took this course, only 122 participated in groups that created and submitted their coursework using CLE. The rest either did not use CLE at all or decided halfway through the assignment to switch from CLE to more traditional methods of collaborating, such as shared Word documents. In total, data of 122 students partitioned into 72 groups across the 3 subsequent assignments were collected, as detailed in Table 1. The group sizes varied typically between 3 and 6 students, however there were instances when only 1 or 2 students contributed within the CLE. The groups were prescribed by the teacher, and for each of the assignments, the students were assigned to new groups. For each assignment, a grade was allocated to the whole group as a result of teacher assessment on the basis of the quality of the joint reports consisting of 9 assessment criteria that are format, abstract, executive summary, introduction and overview, problem statement and problem framing, design alternatives considered, evaluation of alternatives, basis for design selection, results of comparison of the alternatives, and appendices for supporting materials. It should be noted that the grades for the students were unrelated to their CLE collaboration. In the original data, the grades for the first two assignments range within [0,5] and those for the third assignment range within [0,30]. To make the grades comparable across all 3 assignments, they have been normalized within [1,5] for classification and [1,100] for regression. The distributions of the group sizes and normalized assignment grades are as shown in Fig. 5.

The data set used in the experiment is small in terms of the number of examples, and imbalanced with respect to grade distribution as most of the samples received grades of 4 and 5. Despite these limitations we made several provisions to extract the maximum insights and value from these data while trying to maximize the reliability of the generated outcomes and the corresponding conclusions. Specifically, we tried to ensure the features extracted for the predictive models contain maximum discriminative power with respect to the target of prediction. Moreover, given the small data set, we limited the number of features to between 3 and 6 throughout the experiments in order to avoid overparameterization. We also tried to eliminate excessive data imbalance with respect to target classes by fine-tuning the predictive models and modifying the cost functions to better focus on predicting underrepresented classes. Finally, throughout the evaluation we used the 10-fold cross-validation method for assured estimation of the implemented predictive models' performance. In-line with this method we first split our data set into 10 parts. In the subsequent experiments 9/10 parts of

Q23 **Table 1** Description of data collected via CLE platform developed at EBTIC and used in experiments

No. students	122
No. assignments	3
No. groups in 3 assignments	26, 26, 20
Min. size of groups	1
Max. size of groups	6

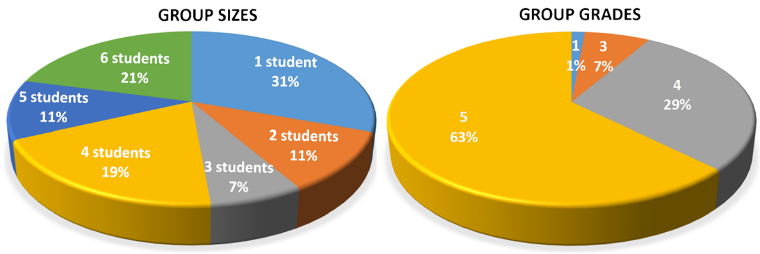


Fig. 5 Distributions of group sizes and group assignment grades

the data were used for building and training the predictive model and the remaining 1/10 part was used for testing the model performance. These experiments were repeated for 10 different splits into training and testing sets and the results aggregated to achieve a reliable, validated accuracy estimate. Note that in such a method the whole data set was exposed for training and testing at different splits and every single point was used at least once as a testing sample. This method is very effective particularly for small data sets and eliminates the risk of dependency on accidental or 'lucky' order of data taken for training and testing.

Group performance prediction

This subsection is dedicated to the experimental results of group performance prediction for completed groupwork tasks based on group interaction data using both classification and regression models. With the classification model, we compared its prediction performance using the ELM and CART as learning algorithms on various feature sets. The CART was also applied as a regression model to predict the groups' scores within [1,100]. The results are elaborated below.

Classification based group performance prediction

To test the classification model for group performance prediction, the whole data were partitioned into training and testing sets. The training set was used to train the prediction model that was then tested using the testing set. First we employed the ELM feedforward neural network for grades prediction. In our implementation the number of hidden neurons was set to be 50 and the sigmoidal function was used as the activation function.

A. Individual features

Following an exploration of various interaction-based features bearing high predictive power for group performance predictions, we concluded with several feature definitions: the number of revision and the length of changes representing the contribution and interaction among students within the group, and the size of the group. These features reflect the amount of knowledge exchange and collaboration available during the learning and content generation processes, all of which can affect group performance (Cen et al. 2014b). To evaluate their individual predictive power we tested them independently using ELM model in 10-fold cross-validation. The average grade prediction accuracies using independently the group size, the number of revision and the length of changes, were respectively: 0.59, 0.61, and 0.63 for training and 0.48, 0.51, and 0.53 for testing, indicating significant predictive power for this 5-class classification problem.

B. Feature fusion

The individual features were then combined together to predict learning performance. Initially, the grades of the groups were predicted using the standard features given in Eq. (3) without considering the diversity of different assignments and students. The average accuracies achieved with the training and testing sets were 0.67 and 0.58, respectively. They are illustrated in the first column group in Fig. 6, where the heights of the boxes filled with green and gray are the average accuracies in training and testing, respectively. It can be seen from the results that the accuracy is quite low without normalization.

The statistics of prediction accuracies achieved with different feature sets across 10-fold validation are compared in Fig. 7, where the 4 boxes in each of the 2 figures illustrate the independent results achieved by using group size, number of revisions, length of changes, and the feature set defined in (3) respectively, in training and testing. From the figure, we can see that by combining the 3 types of features together, we can achieve much better results than by applying them individually.

Next, we considerd the assignment diversity in prediction with the features calculated according to Eq. (7). With the same settings in the ELM, the average accuracies achieved in 10-fold validation with the training and testing sets were 0.69 and 0.62 respectively, and are shown in the second column group in Fig. 6. Compared to the previous model, the testing accuracies improved only slightly yet their stability, measured by the standard deviation over performances from individual cross-validation splits, improved from 0.13 down to 0.05. This illustrates that feature normalization based on assignment diversity can help to improve prediction performance with higher accuracy and better stability.

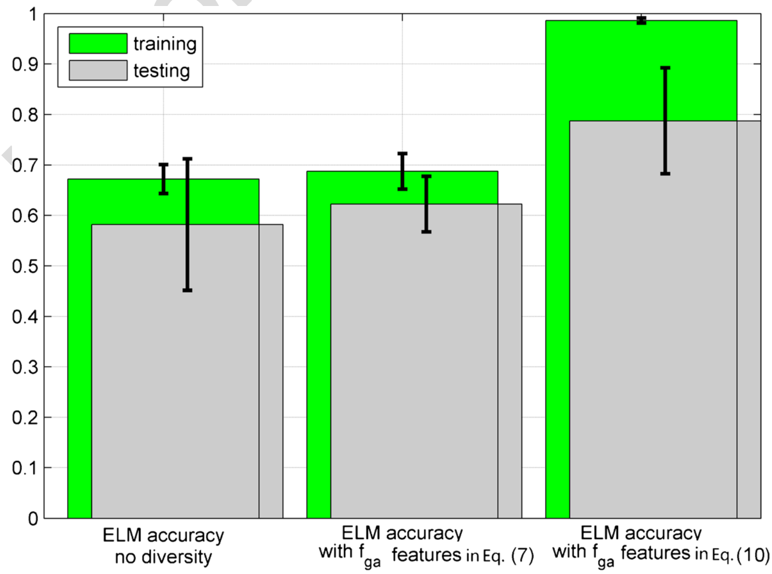


Fig. 6 ELM training and testing prediction accuracies in 10-fold cross-validation using standard features and the features defined in Eqs. cuu, respectively. The corresponding testing accuracies were 0.58, 0.62, and 0.79

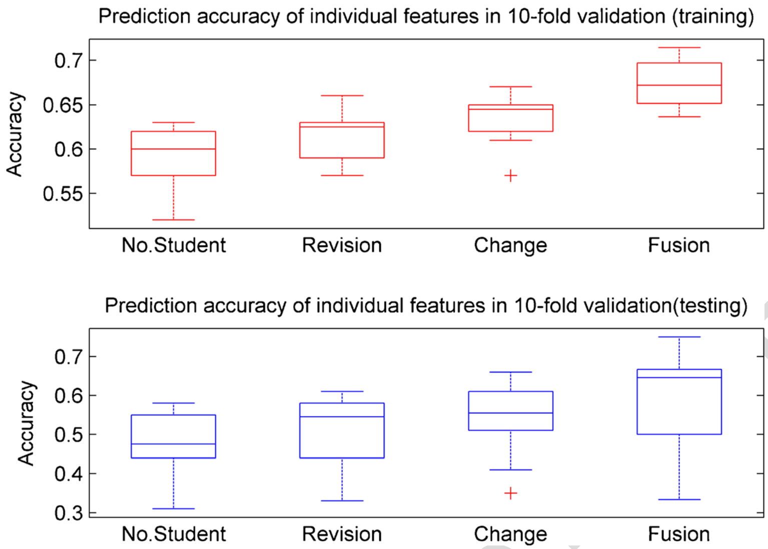


Fig. 7 ELM training and testing prediction accuracies in 10-fold cross-validation by independently using group sizes, the number of revision and the length of changes as features, together with the standard features for comparison

Finally, group performance was predicted by considering the diversity of students with the features calculated according to Eq. (10). The accuracies of training and testing are shown in the third column group in Fig. 6. The average accuracies were 0.99 and 0.79, respectively. The accuracies in both training and testing were largely improved compared to both previous models which have not considered student performance weights in feature generation.

For reference, individual training and testing accuracies obtained for all 10 cross-validation splits are presented in Table 2. The effect of normalization is clearly visible from the table, where in most instances shown in the 3rd group achieved with the normalized features, the predictive accuracies are much higher than in the other 2 groups.

Next, we employed the CART as the learning algorithm in the classification model for performance prediction. The average testing accuracies achieved in 10-fold validation were

Table 2 Training and testing results from 10 cross-validation splits using ELM with different feature sets

	ELM standard		ELM + f_{ga} (7)		ELM + f_{ga} (10)	
Split no	Training	Testing	Training	Testing	Training	Testing
1	0.7143	0.4444	0.6984	0.5556	1.0000	0.6667
2	0.6563	0.5000	0.6563	0.6250	0.9844	0.6250
3	0.6875	0.5000	0.6563	0.6250	0.9844	0.7500
4	0.7031	0.6250	0.7344	0.6250	0.9844	0.7500
5	0.6875	0.7500	0.7344	0.6250	0.9844	0.7500
6	0.6364	0.6667	0.7273	0.6667	0.9848	0.8333
7	0.6515	0.6667	0.6667	0.5000	0.9848	0.8333
8	0.6515	0.6667	0.6515	0.6667	0.9848	0.8333
9	0.6970	0.3333	0.6515	0.6667	0.9848	1.0000
10	0.6364	0.6667	0.6970	0.6667	0.9848	0.8333

0.59, 0.72 and 0.839, with the feature set described in Eqs. (3), (7), and (10) respectively. A more detailed comparison between the performance of ELM and CART, including the prediction accuracies of both training and testing sets, is summarized in Table 3. It confirms the superiority of the diversified feature definitions in Eq. (10) and also indicates a slight edge of the CART model over the ELM one in terms of the accuracy of predictions over testing sets. The results indicate that both the diversity of assignments and students' individual skills should be explicitly factored in the data feature representations when the objective of the task is to predict the performance in the group assignment.

Regression based group performance prediction

In this subsection, the learning performance of groups was predicted using the CART regression tree model. All grades were normalized within the range of [1,100] as discussed in Section 4.

The features were calculated according to Eq. (10). The correlation between the actual and predicted grades of the group and the Root Mean Squared Error (RMSE) achieved for training and testing sets across the 10-fold cross-validation are illustrated in Fig. 8. The average correlation values were 0.94 and 0.82 while the average RMSE were 4.9 and 7.73 for training and testing respectively. The RMSE values are below 10 that is usually considered as a unit in formal assessment, which indicates the applicability of our prediction model in practice.

Figure 9 shows the actual and predicted values of group grades obtained from 10-fold cross-validation for both the learning and testing sets. As can be seen from the figures, its predictions are consistently quite close to the blue lines corresponding to prefect predictions.

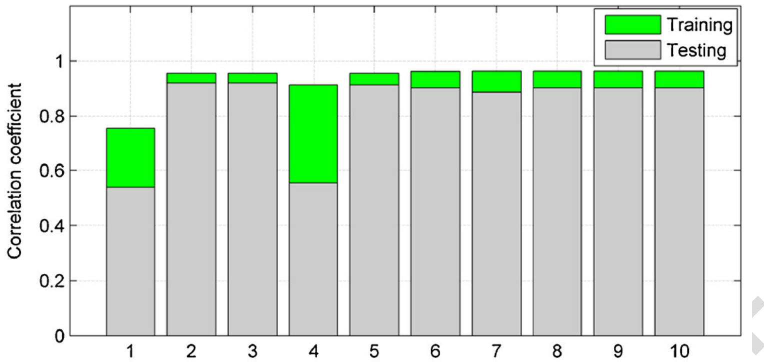
The impact of groupwork on learning outcomes

The aim of this section is to investigate the impact of collaborative groupwork on student learning outcome. We intended to quantitatively establish whether or not student performance could be improved through collaborative learning. The student performance evaluation was carried out using the comparative analysis model presented in Section 5.

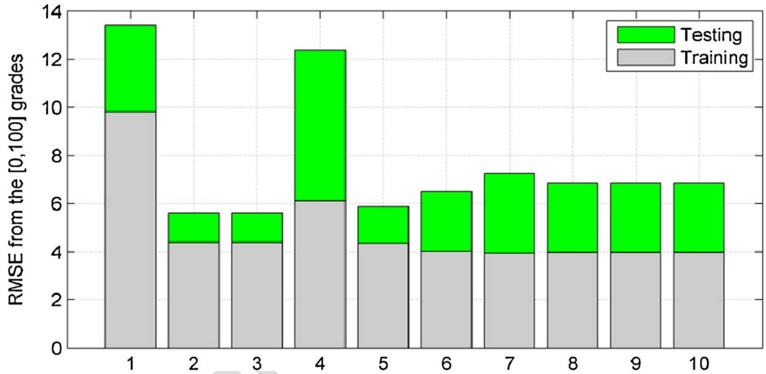
First, we compared the students' estimated grades with the grades actually received by the groups to which the students belonged. The grade deviation was taken as an indicator as to whether or not the students' performance could be improved through collaborative learning. Among 239 student-assignment instances, 122 had higher group performance than their individual student performance expectations, 40 had lower group performance, and 77 had the same group performance, all of which are shown in Fig. 10(a). It can be seen from the figure that the group performance of most of the students is higher than their individual performance expectations. This indicates that collaborative learning results in additional

Table 3 Comparison of ELM and CART accuracies from training and testing with different feature sets

	ELM			CART		
Feature Eq.	(3)	(7)	(10)	(3)	(7)	(10)
Training	0.67	0.69	0.99	0.70	0.79	0.844
Testing	0.58	0.62	0.79	0.59	0.72	0.839



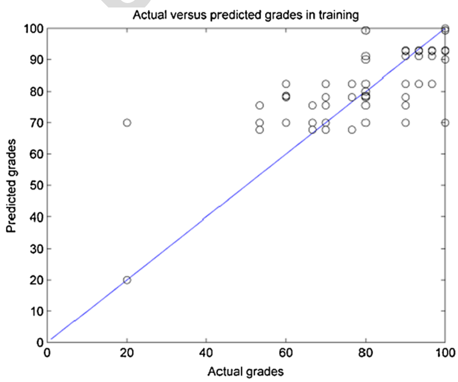
(a) CART regression correlation in 10-fold cross-validation



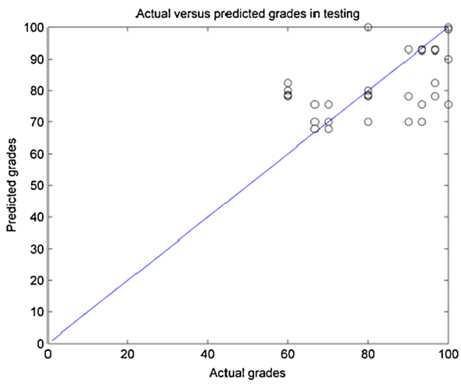
(b) CART regression RMSE in 10-fold cross-validation

Fig. 8 CART regression correlation between actual and predicted group grades and Root Mean Squared Error over 10 cross-validations splits for training and testing

learning synergy that positively impacts on their performance in the group. Thus, a better learning outcome can be expected when learning in a group as compared to learning alone.



(a) Training set results



(b) Testing set results

Fig. 9 Actual vs predicted grades obtained with CART model via 10-fold cross-validation

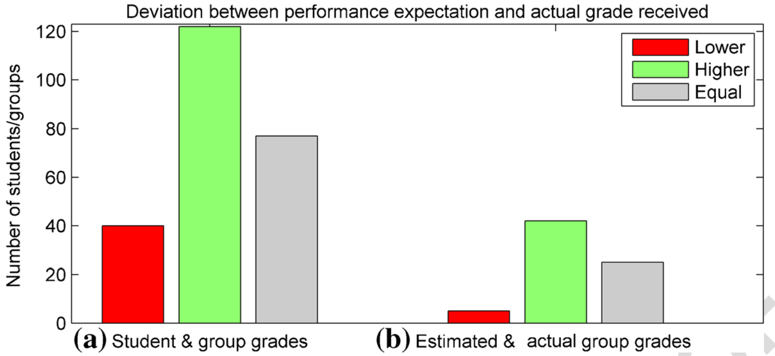


Fig. 10 Number of students who have lower, higher and equal group marks compared to their estimated individual marks; and number of groups that have lower, higher and equal group marks compared to their estimated group marks

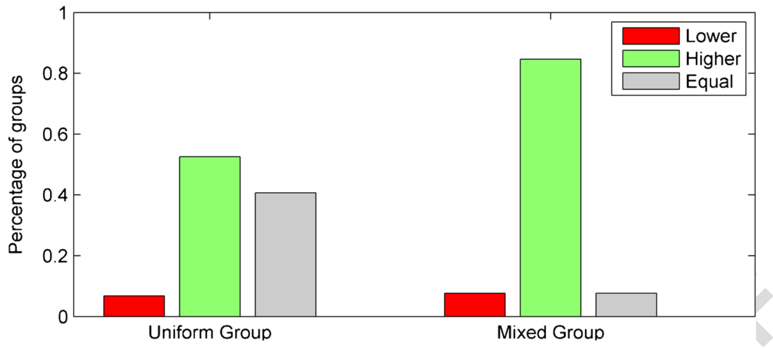
Second, we modeled the grade of a group as a linear combination of contribution-weighted performances of all individual students in the group according to Eq. (22). The deviation between the group performance expectation and the actual grade received is likely to be linked solely to students collaboration in the group and therefore can be considered as a measure of collaborative learning synergy. Collaborative learning synergy is qualitatively defined as absorbing knowledge and creation of educational content with performance exceeding students prior performance expectations. This is the case where the quality of the creative content from a group of students appears to exceed the sum of their expected contributions due to the value-added effects of stimulation, mutual reflection, dynamic exploration, meaning-making and continuous feedback. This is in agreement with the intellectual synergy of many minds working on a problem and the social stimulation of mutual engagement in a common endeavor produced by collaborative learning reported by Golub (1988). From the quantitative point of view, we defined the synergy of group collaboration simply as a difference between the actual group assessment and the expected group assessment understood as an average of the performances of group members. It is considered to be a much simpler and clearer definition which is easily measurable in the experiments as opposed to the qualitative definitions which are hard to quantify.

Overall there were 72 group-assignment instances, each of which represented one group assigned to one assignment. Among them, the actual marks in 42 groups were higher than the estimates, 25 groups had equal estimated and actual marks, and only 5 groups reported lower actual marks, which are shown in Fig. 10(b). The observed collaborative learning synergy is quite pronounced here: 58.3 % of the groups had the actual group performance better than the sum of their individual student performance contributions, while only 6.9 % of the groups had their actual performance below their expectation.

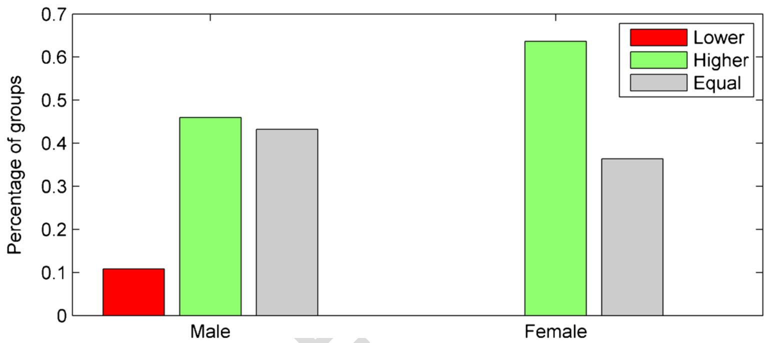
The impact of group composition

Gender composition

Among the 72 group-assignment instances, there were 59 uniform-gender groups and 13 mixed-gender groups. Figure 11(a) shows a comparison between the actual and



(a) Uniform vs mixed genders comparison



(b) Male-only vs female-only groups comparison

Fig. 11 Comparison of individual expectations and actual group grades

estimated marks for both of the uniform and mixed groups. It can be seen that 52.5 % of the uniform groups had their actual marks higher than the estimated ones, while 84.62 % of the mixed groups achieved higher marks than their estimated ones. This indicates that co-education with mixed-gender groups could further stimulate groupwork synergy and push the improvement of their learning performance to even higher levels. Although the advantages of co-education have been pointed out in the literature, e.g. (Cen et al. 2014a; Crosswell, and Hunter 2012; Smith 1996), we have shown via measurable quantitative analysis the way to verify and quantify such phenomenon, and we have also provided the methodology for assessing and estimating how much performance improvement can be driven by and attributed to co-education. For the uniform-gender groups, it has been observed that the female-only groups performed better than the male-only groups in this particular course, which is shown in Fig. 11(b).

Ability level composition

Based on the proposed individual student performance estimates, all students were categorized into 5 levels: [0–70), [70–80), [80–90), [90–100), 100. The ability distribution of 239 student-assignment instances are illustrated in Fig. 12.

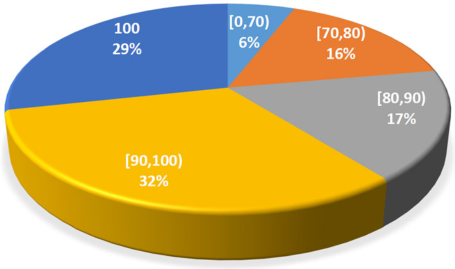


Fig. 12 Ability (grades) distribution among students-assignment instances

Figure 13 illustrates the impact of collaborative learning on group performance for students with different ability levels. It can be seen that more than 70 % of lower-ability students with grades lower than 80 had better group grades than their individual performance. Also for around 90 % of high-ability students with grades in the range [90,100] their performance was also improved via collaboration. The same level of improvement was likely valid for top students with 100 % scores as well but is somewhat hidden in the "Equal" category. Interestingly these results indicate that top- and bottom-ability students are more likely to improve their performance through collaboration in heterogeneous groups consisting of students with different ability-levels. The highest-ability students, however, were not largely influenced by group composition when they were working with students having relatively lower abilities. This quantitatively proves and is consistent with the finding presented in (Webb et al. 1998). It is also interesting to note that more than 60 % of medium-ability students with grades within [80–90) achieved lower group grades in collaborative learning setup. This could be due to the observation that in heterogeneous groups with wider ability range, higher-ability and lower-ability students tend to form teacher-student relationship with more interaction and collaboration, while medium-ability students tend to be left out and participate less, as also reported by Webb (1991).

The above analysis was devoted to student performance modeling and understanding the impacts of various factors and characteristics of group composition and the mechanics of group interaction and groupwork generation in the context of collaborative learning. The results evidentially suggested that diverse groups with diversity of skills, abilities and even mix of genders are more likely to benefit from the synergy generated in collaborative learning and hence achieve much better learning outcomes compared to just individual learning alone.

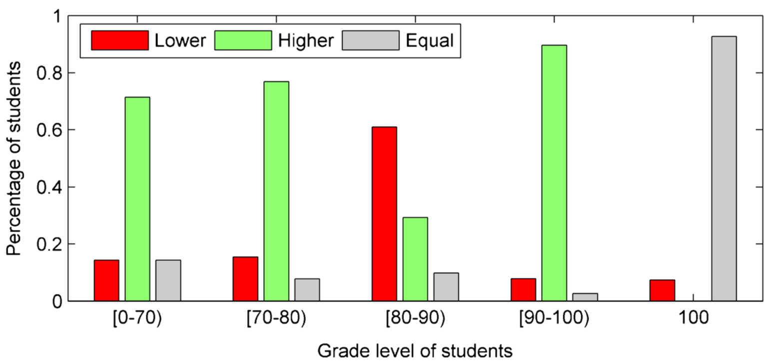


Fig. 13 The impact of collaborative learning on performance of students with different ability levels

Discussion

The experimental results described and reported above have illustrated the effectiveness of the proposed quantitative approaches to measurement, prediction and impact analysis in computer-supported collaborative learning. Specifically, we have shown how predictive models equipped with supervised learning engines can be successfully used for group performance prediction based on different amounts of evidence at different stages of the group exercises (although we have only shown the results of predictions made after completion of the groupwork courseworks i.e. utilizing the complete interaction evidence from the groupwork activities). We pointed out that while prior individual student performance estimates usually provide a good estimate of likely group performance, the way a group collaborates and its individual members interact is crucial to generate additional collaboration synergy benefits, though these are by no means guaranteed. Live group performance prediction based on interaction data automatically collected by the collaboration-enabling system, offers a whole new layer of benefits ranging from much more reliable group performance estimates, through monitoring individual students' contribution to the groupwork, up to early identification of under-performing students and communicating the high risk of failure to both the teachers and affected students for appropriate corrective actions. Although we have not presented detailed reconciliation between the predicted and actual groupwork performance, beyond just the average statistics, such comparison can lead to discovery of inconsistencies in teacher assessment; this could become apparent when the deviation between data-based predictions and the actual assessment is significantly above the range of synergy impact and perhaps in conflict with the collaboration-activity data.

The methodology we proposed for group performance prediction can be easily integrated into a real-time system for automated and continuous expectation of student educational performance; this would allow the student to make more informed decisions about his/her curriculum and career path choices throughout the curriculum. It can provide real-time performance prediction from the beginning to the end of learning process depending on previous student experiences and live interaction data if available. It can extend the prediction of possible group performance to what-if scenarios before the group has even formed, and with this respect could be used as a criterion for optimized automated group formation. Such a groupwork performance-driven predictive recommendation engine could be an asset in every academic institution that would ensure the full exploitation of individual students' potential and more efficient utilization of students' and teachers' time, with the ultimate goal of turning education into an enjoyable and satisfiable experience with maximum knowledge transferred and retained among the students.

A comparative student performance model has been proposed to assess the performance of individual group members, which allows teachers to quantitatively analyze the learning qualities of individual students based on their contributions to completed joint assignments and group achievements. In addition, a generative mixture model has been proposed to isolate the impact of collaboration style from the individual student qualities on group performance. Based on this model, various forms of group composition are quantitatively analyzed, and some useful grouping rules, which are either supported or disputed in the literature, are suggested and quantitatively assessed.

In our method, students' interaction and contribution are quantified using the number of revisions and the length of changes performed in collaborative writing. As shown in the experimental results, it can work quite well as with our data collected across 3 collaborative writing tasks. It should be noted that for some courses, e.g. mathematics, which requires fewer text inputs to complete assignments and where small changes may lead to totally different answers and, consequently, quite different scores, our method may not yield good results. Text analysis to understand context and content can be more helpful, although it will make solutions more complex in terms of computational cost and system implementation. In subsequent work, we intend to extend this research by exploiting more evidence in the form of student profiles, complete journey through out their curricula and the actual content of assessed assignments. Multimedia components, like spoken-dialogue of discussion, will also be analyzed to catch interactive activities during the learning process.

In this work, the prediction models are based on machine learning techniques. Although some of the data-driven ML models could be difficult to interpret, the models we utilized are far from black boxes. Both ELM and CART models directly express the relationship between the input characteristics of collaborative learning and the formal learning performance. ELM is in a version of neural network models which can be visualized to gain insight into how the outputs (class supports) are formed from the interconnected weighted links rooted from the inputs. In turn CART, as an instance of decision tree, is one of the most transparent models that can be shown as a tree of conditions upon the features directing the decision along the feature-branches to the leaves (decisions). The decision tree model can be fully converted into SQL code which is nothing more than a stream of if-then conditions applied to the raw data, and hence is extremely easy and explicit to work with. Subsequent work will investigate the approaches to further improve the comprehensibility of the models so that instructors can accurately measure to which extent individual factors affect learning performance. We also intend to expand the predictive span of our systems into delivering predictive performance-driven recommendations on modules, courses and/or knowledge contents that each individual student likely to be best at, and hence fulfilling his/her educational and career goals with satisfaction and accomplishment. In addition, research and development will be extended to the live scenario of utilizing incomplete groupwork interaction data to attain real-time applicability of the presented methodology in a classrom environment, for instance, to dynamically re-organize students in groups with poor expected performance predicted during the learning process or at the early stages of joint educational activity. Such non-trivial attempts would make a significant step forward to make the performance prediction models more applicable in practice.

The quantitative approaches proposed in our work use simple general features to represent contributions and interactions among students, e.g. types and amounts of text editing in collaborative writing tasks. As such, they can be applied not only with the data collected using our CLE platform, but also with the collaboration data generated by other CSCL platforms. As mentioned before, the dataset used here has a very limited number of examples and is imbalanced with respect to the target grade classes. However, the results are still reliable based on the following consideration. First, in group performance prediction, distinctive features are extracted to represent the contributions and interactions among students. Second, the feature sets are limited to avoid overparameterization, while the feature definitions themselves are normalized to allow comparable utilization across diverse instances of group exercises, skills of group members, and group assessments. Third, throughout the experiments we used 10-fold cross-validation as a reliable performance estimation method which is especially suitable for small datasets. We have shown that the proposed methods are reliable and stable with acceptable accuracy and small standard variation. Fourth, in the comparison between actual group grades and predicted individual performance, 122 among 239 student-assignment instances show group performance higher than individual student performance expectations, while only 40 have lower group performance. It can be seen that the number of students whose group performance is higher than their individual performance expectation is 3 times of those with lower group performance. Similarly, 42 of the 72 group-assignment instances have higher actual marks than their estimated ones, and only 5 groups have lower actual marks. Although there are a limited number of instances to calculate the statistics, the difference between the two counterparts is high enough to make a significance claim related to the benefits of collaborative learning and its resulted synergy. Finally, in the comparison between the single- and mixed-gender groups, there are 52.5 % of 59 instances with uniform groups having their actual marks higher than the estimated ones, while 84.6 % among 13 instances with mixed groups achieving higher grades than their prior estimates. The large difference between the two indicates that the results are rather credible and convey valid conclusions despite the small data size and class imbalance. In future work, an extended dataset with more diverse data, for instance, a dataset with more groups containing more active students and a bigger spread of grades or teacher assessments, will be collected and intensive experiments will be conducted for further evaluation and formal validation (e.g. evaluating the improved prediction models with f-measure and statistical tests).

Group composition is quite important in collaborative learning, which may affect group learning performance considerably. Automatic group formation approaches based on global optimization and clustering will be explored in subsequent work utilizing the evaluation criteria's key drivers identified in this work. The comparative student performance assessment model will be further validated and matched using standard reference systems like the students' actual Grade Point Average (GPA).

Conclusions

In this work we have made a pioneering effort to quantitatively describe the characteristics of collaborative learning and assess their impact on group academic performance. We wanted to convey a generic message that data-driven prediction of group performance could be an effective criterion not only to gain an immense, objective and quantitative insight into how and why collaboration is effective for learning, but also to hint at how it can guide the whole start-to-end process of group learning from group composition, through live group interaction monitoring to post-assessment consistency analysis and performance-driven recommendations.

We first focused on the central problem of predicting group performance which can be considered as the enabler of our methodology. We have shown that machine learning and, in general, predictive analytics are now mature enough to provide reliable predictions of group performance at every stage of group exercise: before, during and after its completion. We have shown that while individual prior performances are good estimate of expected group performance, live group interaction data offer much richer evidence that can lead to more reliable predictions of group performance that takes into account its resulted collaboration synergy. We used both classification and regression models to predict group performance based on students' interaction data extracted from the trial of the Collaborative Learning Environment (CLE) platform developed at EBTIC. We defined a set of discriminative features from group sessions of concurrent student learning and interaction sequences as they were working on the group coursework. These features measured various characteristics of individual members' interactions and contributions to the joint assignments, and were designed to differentiate between different outcomes represented by the formal group assessment. The challenges posed by the necessity to accommodate different students, diverse assignments and assessment methods have also been addressed and resolved through normalized and unified assessment representations and generic normalized feature definitions.

Extreme Learning Machine (ELM) based feedforward Neural Networks (NN) and Classification and Regression Trees (CART) were used as representative instances of Machine Learning techniques applied to predict group performance in accordance with the features derived from group interaction data. The series of experiments have been carried out on the data collected from the CLE trial that ran with 122 students from the courses of Molecular Biology Engineering and the Freshman Design Engineering at Khalifa University. The results revealed many interesting insights. The accuracy of group's grade predictions in the classification setup was in excess of 80 %, while the CART model, set up in the regression mode, reported an error rate of below 8 %. These are rather impressive results suggesting that just based on the timely style and intensity of collaborative learning we seem to be capable of predicting group grades with an average error of less than one or half a grade, respectively. These prediction results were obtained after observing the complete evidence extracted from the groupwork. However, the exact methodology can also be extended, without any loss of generality, into a live scenario of real-time groupwork performance prediction with limited expected performance losses and quick convergence to the final stable predictions, although the detailed analysis of such real-time framework still remains the subject of our future work.

This fairly good group performance prediction capability was then back-propagated and decomposed to provide explanations into how, when and why collaborative learning really works. To capture the essence of the collaboration, we have developed a comparative performance model to evaluate the academic value of individual students in relation to its group performance. This is quite useful for the teachers to understand the hidden performance of individual students in collaborative learning where otherwise the assessment would be overlooked based on the achievement of the whole group. This model then evolved and was improved through a decomposition utilizing generative mixture of group performance. It assumes that the group assignment grade is generated as a combination of fixed students grade expectations, improved or degraded by the collaboration type that the students choose to follow. Both models provide new interesting ways to quantitatively analyze the improvement or degradation achieved through collaborative learning exercises. We have shown, via numerical analysis, that the students indeed do improve their academic performance through learning in groups compared with their individual performance expectations.

What differentiates our work from the others in the field, however, is that we have proposed a simple and measurable quantitative definition of collaboration synergy that directly measures the deviation between the average individual performance expectation and the actual group assessment. Such defined synergy is an isolated measure of the quality of collaboration that solely determines whether the students will benefit or lose out from collaboration following specific patterns of interaction and groupwork. The beauty of this approach is that such defined synergy can itself be a subject of prediction and the data features that provide the most explanation can thereby be identified as key drivers of synergy in group collaboration. Our experimental results clearly indicate that higher synergy is obtained in groups with a high diversity of skills, equal-distribution of workload and high concurrency of interaction with as many members as possible.

Finally our work concludes with the groundwork for quantitative analysis on the impact of group composition on learning performance. This revealed several very interesting findings and hints for future promising research directions. Specifically, experiments with synergy predictions back-propagated onto group composition characteristics, especially skill and gender distribution in the group, provided numerical evidence to support the claim that gender diversity in the group and, separately, the diversity of student skills or abilities do improve the group performance. Members of the group with mixed gender are observed to engage, contribute and perform significantly better compared to the uniform-gender groups. Additionally, the groups with a mixture of low and high performing students tend to benefit the most from groupwork, due to the apparent emerging student-teacher relationships, which stimulate students' engagement, knowledge exchange and reflection on mutual input; while medium-performing students appear to be a bit left out and participate less. Backed by the retrospective reflection, this intriguing observation was explained by the emergent tendency to form micro-subgroups or pairs within the groups that take over the communication channel in the group. Although such self-organizing sub-clustering is in general a very desirable property of group interaction, it remains open to see if this can be further utilized to better distribute collaboration benefits among all the members of the group and how that can be further encouraged.

References Q24

Araque, F., Roldan, C., & Salguero, A. (2009). Factors influencing university drop out rates. *Computer and Education, 53*, 563–574.

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: a review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17.

Barber, R., Sharkey, M. (2012). Course correction: using analytics to predict course success. *Proceedings of the Second International Conference on Learning Analytics and Knowledge, ACM*

Barkley, E. F., Cross, K. P., Major, C. H. (2004). *Collaborative learning techniques: A handbook for college faculty*. Jossey-Bass. Q25

Bhardwaj, B.K., Pal, S. (2011). Data mining: a prediction for performance improvement using classification. *International Journal of Computer Science and Information Security*, *9*(4). Q26

Bishop, C. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey: Wadsworth, Inc.

Bruckman, A., Jensen, C., & Debonte, A. (2002). Gender and programming achievement in a CSCL environment. In K. Bruffee (Ed.), *Collaborative learning*. Baltimore: The Johns Hopkins University Press.

Cen, H., Koedinger, K., Junker, B. (2006). Learning factors analysis - a general method for cognitive model evaluation and improvement. *International Conference on Intelligent Tutoring Systems*, Q27

Cen, L., Ruta, D., Powell, L., Ng, J. (2014). Learning alone or in a group - an empirical case study of the collaborative learning patterns and their impact on student grades. *International Conference on Interactive Collaborative Learning*. Q28

Cen, L., Ruta, D., Powell, L., Ng, J. (2014). Does gender matter for collaborative learning? *International iCampus Forum 2014 on Smart Education for the 21st Century, IEEE International Conference on Teaching, Assessment, and Learning for Engineering*, New Zealand.

Changa, G. K., Chenb, G. D., & Wangb, C. Y. (2011). Statistical model for predicting roles and effects in learning community. *Behaviour and Information Technology, 30*(1), 101–111. Q29

Chennabathni, R., Rgskind, G. (1998). Gender issues in collaborative learning. *Canadian Women Studies*, *17*(4). Q30

Chiu, M. M. (2000). Group problem solving processes: social interactions and individual actions. *Journal for the Theory of Social Behaviour, 30*(1), 27–50.

Chiu, M. M. (2008). Flowing toward correct contributions during groups' mathematics problem solving: a statistical discourse analysis. *Journal of the Learning Sciences, 17*(3), 415–463.

Coffrin, C., Corrin, L., Barba, P., Kennedy, G. (2014). Visualizing patterns of student engagement and performance in MOOCs. *ACM Press*, 83–92. Q31

Cohen, E. G., Lotan, R. A., Abram, P. L., Scarloss, B. A., & Schultz, S. E. (2002). Can groups learn? *Teachers College Record, 104*(6), 1045–1068.

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

Crosswell, L., & Hunter, L. (2012). Navigating the muddy waters of the research into single sex class-rooms in co-educational middle years settings. *Australian Journal of Middle Schooling, 12*(2), 16–27.

Davidson, J. E., Sternberg, R. J. (Eds.) (2003). *The psychology of problem solving*. Cambridge University Press. Q32

Dillenbourg, P. (1999). Collaborative learning: Cognitive and computational approaches. *Advances in learning and instruction series*. New York, NY: Elsevier Science, Inc. Q33

Dillenbourg, P. (2000). What do you mean by 'collaborative learning'? In P. Dillenbourg (Ed.), *Collaborative-learning: Cognitive and computational approaches* (p. 119). Oxford: Elsevier.

Dillenbourg, P., Baker, M., Blaye, A., OMalley, C. (1996). The evolution of research on collaborative learning. In E. Spada, P. Reiman (Eds.), *Learning in humans and machine: Towards an interdisciplinary learning science*, (189–211). Oxford: Elsevier. *Proceedings of Computer Support for Collaborative Learning*, 119–127.

Dirkx, J. M., Smith, R. O. (2013). Online collaborative learning. In T. S. Roberts (Ed.), *IGI Global*. Q34

Fall, R., Webb, N., Chudowsky, N. (1997). Group discussion and large-scale language arts assessment: Effects on students comprehension. *CSE technical report*, Los Angeles: Cresst. Q35

Feichtner, S. B., & Davis, E. A. (1991). Why some groups fail: a survey of students experiences with learning groups. *The Organizational Behavior Teaching Review, 9*(4), 75–88.

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction, 19*(3), 243–266.

Ferguson, R., Shum, S.M. (2011). Learning analytics to identify exploratory dialogue within synchronous text chat. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, *ACM*, 99–103.

Goggins, S., Xing, W., Chen, X., Chen, B., & Wadholm, B. (2015). Learning analytics at "small" scale: exploring a complexity-grounded model for assessment automation. *Journal of Universal Computer Science, 21*(1), 66–92.

Goldstein, J., Puntambekar, S. (2004). The brink of change: gender in technology-rich collaborative learning environments. *Journal of Science Education and Technology*, *13*(4). Q36

Golub, J. (Ed.). (1988). *Focus on collaborative learning*. Urbana: National Council of Teachers of English.

Goode, W., & Caicedo, G. (2014). Online collaboration: individual involvement used to predict team performance. *Learning and Collaboration Technologies, Technology-Rich Environments for Learning and Collaboration, Lecture Notes in Computer Science, 8524*, 408–416.

Gordon, A. (2000). In a class of their own: boys benefit even more than girls from single-sex schools, a-level grades study reveals. *The Mail on Sunday (UK)*, *42*. Q37

Gress, C. L. Z., Fior, M., Hadwin, A. F., & Winne, P. H. (2010). Measurement and assessment in computer-supported collaborative learning. *Computers in Human Behavior, 26*(5), 806–814.

Gunnarsson, B. L., Alterman, R. (2012). Predicting failure: a case study in co-blogging. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, ACM.*

Hackman, J. R., & Morris, C. G. (1975). Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (p. 8). New York: Academic.

Hämäläinen, W., & Vinni, M. (2011). *Classifiers for educational data mining*. London: Chapman and Hall/CRC.

Hirsch, B., Hitt, G. W., Powell, L., Khalaf, K., Balawi, S. (2013). Collaborative learning in action. *Proceedings of the IEEE International Conference on Teaching, Assessment and Learning for Engineering*, Bali, Indonesia.

Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006). Extreme learning machine: theory and applications. *Neurocomputing, 70*, 489–501.

Johnson, D.W., Johnson, R.T. (1988). Two heads learn better than one. *Transforming Education, 34*, Q38

Johnson, D. W., Johnson, R. T. (1998). Cooperative learning and social interdependence theory. Retrieved from http://www.co-operation.org/pages/SIT.html.

Johnson, D. W., Johnson, R. T., & Smith, K. A. (1991). *Cooperative learning: Increasing college faculty instructional productivity*. Washington: The George Washington University, School of Education and Human Development.

Johnson, D. W., Johnson, R. T., & Stanne, M. E. (2000). *Cooperative learning methods: a meta analysis*. Minnesota: University of Minneapolis.

Kagan, S. (1994). *Cooperative learning*. San Clemente: Kagan Cooperative Publishing.

Kapur, M., Kinzer, C.K. (2009). Productive failure in CSCL groups. *International Journal of Computer-Supported Collaborative Learning, 4*(1). Q39

Kapur, M., Voiklis, J., Kinzer, C.K. (2011). A complexity-grounded model for the emergence of convergence in CSCL groups. *Analyzing interactions in CSCL* (pp. 3–23). US: Springer. Q40

Koschmann, T. (2002). Deweys contribution to the foundations of CSCL research. *Proceedings of Computer Supported Collaborative Learning*, 17–22. Q41

Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems, 23*(6), 529–535.

Lai, E. R. (2011). Collaboration: A literature review research report. Retrieved from: http://www.pearsonassessments.com/.

Mael, F., Alonso, A., Gibson, D., Rogers, K., & Smith, M. (2005). *Single-sex versus coeducational schooling: A systematic review*. Washington: American institutes for research.

McNely, B.J., Gestwicki, P., Hill, J.H., Parli-Horne, P., Johnson, E. (2012). Learning analytics for collaborative writing: A prototype and case study. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 222–225.

Mitnik, R., Recabarren, M., Nussbaum, M., & Soto, A. (2009). Collaborative robotic instruction: a graph teaching experience. *Computers and Education, 53*(2), 330–342.

Morse, S. (1998). *Separated by sex: A critical look at single-sex education for girls*. Washington: American Association of University Women Educational Foundation.

Oakley, B., Felder, R. M., Brent, R., & Elhajj, I. (2004). Turning student groups into effective teams. *Journal of Student Centred Learning, 2*(1), 9–34.

Race, P. (2001). A briefing on self, peer, and group assessment. *American Educational Research Journal*, assessment series no. 9, York, UK. Q42

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics Part C: Applications and Reviews, 40*(6), 601–618.

Romero, C., Ventura, S., Espejo, P.G., & Hervs, C. (2008). Data mining algorithms to classify students. *Proceedings of the 1st International Conference on Educational Data Mining*, 8–17.

Romero, C., López, M., Luna, J., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers and Education, 68*, 458–472.

Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem-solving. In C. E. O'Malley (Ed.), *Computer-supported collaborative learning* (pp. 69–97). Berlin: Springer.

Ruta, D., Powell, L., Wang, D., Hirsch, B., Ng, J. (2013). Self-organising P2P learning for 21C education. *International Symposium on Smart Learning for the Next Generation*. Q43

Safin, S., Verschuere, A., Burkhardt, J., Dtienne, F., & Hbert, A. M. (2010). Quality of collaboration in a distant collaborative architectural educational setting. *International Reports on Socio-Informatics, 7*(1), 40–48.

Saner, H., McCaffrey, D., Stecher, B., Klein, S., & Bell, R. (1994). The effects of working in pairs in science performance assessments. *Educational Assessment, 2*(4), 325–338.

Savicki, V., Kelley, M., & Lingenfelter, D. (1996). Gender, group composition, and task type in small task groups using computer-mediated comunication. *Computers in Human Behavior, 12*, 549–565.

Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., Wani, E. (2011). Prediction of student academic performance by an application of data mining techniques. *Proceedings of International Conference on Management and Artificial Intelligence, 6,* 110–114, Bali, Indonesia.

Slavin, R. (1990). *Cooperative learning: Theory, research and practice*. New Jersey: Prentice Hall.

Slavin, R., & Cooper, R. (1999). Improving intergroup relations: lessons learned from cooperative learning programs. *Journal of Social Issues, 55*(4), 647–663.

Smith, I. (1996). Good for boys and bad for girls? Empirical evidence on the coeducation/single-sex schooling debate. *Forum of Education, 51*(2), 44–51.

Stahl, G., Koschmann, T., & Suthers, D. (2006). Computer-supported collaborative learning: A historical perspective. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 409–426). Cambridge: Cambridge University Press.

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

Stump, G. S., Hilpert, J. C., Husman, J., Chung, W. T., & Kim, W. (2011). Collaborative learning in engineering students: gender and achievement. *Journal of Engineering Education, 100*(3), 475–497.

Suthers, D. D. (2006). Technology affordances for intersubjective meaning-making: a research agenda for CSCL. *International Journal of Computer Supported Collaborative Learning, 1*(2), 315–337.

Thai-Nghe, N., Janecek, P., Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. *Proceedings of the 37th IEEE Frontiers in Education Conference*, T2G7–T2G12.

Thai-Nghe, N., Busche, A., Schmidt-Thieme, L. (2009). Improving academic performance prediction by dealing with class imbalance. *Proceedings of 9th IEEE International Conference on Intelligent Systems Design and Applications*, 878–883.

Thai-Nghe, Drumond, N.L., Krohn-Grimberghe, A., Schmidt-Thieme, L. (2010). Recommender system for predicting student performance. *Proceedings of the 1st workshop on Recommender Systems for Technology Enhanced Learning, 1*, 2811–2819.

Thai-Nghe, N., et al. (2011). Multi-relational factorization models for predicting student performance. *KDD 2011 Workshop on Knowledge Discovery in Educational Data*.

Thai-Nghe, N., Horvath, N., Schmidt-Thieme, L. (2011). Factorization models for forecasting student performance. *Proceedings of the 4th International Conference on Educational Data Mining*.

Van Boxtel, C., Van der Linden, J., & Kanselaar, G. (2000). Collaborative learning tasks and the elaboration of conceptual knowledge. *Learning and Instruction, 10*(4), 311–330.

Vita, G. D. (2005). Fostering intercultural learning through multicultural group work. In J. Carroll & J. Ryan (Eds.), *Teaching international students: Improving learning for all* (pp. 75–83). Abingdon: Routledge.

Ward, C. (2006). *International students: Interpersonal, institutional and community impacts*. Wellington: New Zealand Ministry of Education.

Webb, N. M. (1991). Task-related verbal interaction and mathematical learning in small groups. *Research in Mathematics Education, 22*(5), 366–389.

Webb, N. M. (1993). Collaborative group versus individual assessment in mathematics: processes and outcomes. *Educational Assessment, 1*(2), 131–152.

Webb, N. M. (1995). Group collaboration in assessment: multiple objectives, processes, and outcomes. *Educational Evaluation and Policy Analysis, 17*(2), 239–261.

Webb, N. M., Nemer, K. M., & Chizhik, A. W. (1998). Equity issues in collaborative group assessment: group composition and performance. *American Educational Research Journal, 35*(4), 607–651.

Wolff, A., et al. (2013). Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment. *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, *ACM*.

Xing, W., Wadholm, B., Goggins, S. (2014). Learning analytics in CSCL with a focus on assessment: an exploratory study of activity theory-informed cluster analysis. *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*, 59–67.

Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable genetic programming: integrating learning analytics, educational data mining and theory. *Computers in Human Behavior, Elsevier, 47*, 168–181.

Yadav, S. K., & Pal, S. (2012). Data mining: a prediction for performance improvement of engineering students using classification. *World of Computer Science and Information Technology Journal, 2*(2), 51–56.

Zeid, A., El-Bahey, R. (2011). Impact of introducing single-gender classrooms in higher education on student achievement levels- a case study in software engineering courses in the GCC region. *Proceedings of the 41st ASEE/IEEE Frontiers in Education Conference*.

Zheng, L., Huang, R. (2016). The effects of sentiments and co-regulation on group performance in computer supported collaborative learning. *The internet and higher education*, 28, 59–67, Elsevier.

Zhu, C. (2012). Student satisfaction, performance, and knowledge construction in online collaborative learning. *Journal of Educational Technology and Society, 15*(1), 127–136.