http://dx.doi.org/10.14236/ewic/HCI2022.12

Designing a Peer Assessment for Identification of Students' Group Work Problems

Danny Veth
Utrecht University, the Netherlands
d.veth@students.uu.nl

Isabella Saccardi
Utrecht University, the Netherlands
i.saccardi@uu.nl

Judith Masthoff
Utrecht University, the Netherlands
j.f.m.masthoff@uu.nl

Group projects are part of the core educational experience in higher education, but many students report bad experiences. Group problems may undermine learning and cause stress and frustration. This may be prevented by monitoring and supporting groups, but this is often not feasible for teachers, who lack time and resources. This research aims to find a method for early identification of group work problems via computer-supported assessment. First, interviews and focus groups provided insights into the most common group problems and which visual features students preferred in a peer assessment. Next, two assessment versions were created: a simple, time-efficient version, and a more engaging, interactive one. We also created an initial version of E-Mate, a virtual agent that provides initial feedback on the assessment. These were tested in a field study. Most students reported a positive experience with the peer assessment, regardless of the visualization used. Teachers were also positive about its usefulness. The research also supports the use of five attributes to assess group collaboration.

teamwork, group problems, survey design, peer assessment, emotional support

1. INTRODUCTION AND RELATED WORK

Group work is increasingly popular in higher education. Group projects are known to encourage the acquisition of interpersonal and communication skills, which are highly valued in many professional fields (Colbeck et al. 2000). Unfortunately, they do not always run smoothly: conflicts can arise and cause stress and frustration (Burdett 2003). Identifying such problems in an early stage can have a large positive impact on both students and teachers. Early detection can prevent escalation, manage frictions and encourage positive communication from early on (Adeniran 2020). However, it is often unrealistic for teachers to closely monitor a large number of groups to mitigate and detect group conflicts: the available time is scarce and the number of students in a given class may exceed their attention. Software systems can play an essential role in the early identification of conflicts by supporting group monitoring and making issues known to teaching staff (Freeman and McKenzie 2002; Murray and Boyd 2015; Badea and Popescu 2019). Our research aims to create a system that assesses groups to detect issues early, provide some support and signal the necessity of an intervention to teachers, diminishing their workload while promoting groups' well-being.

1.1. Group problems

Group problems may not only undermine learning goals, but also affect individuals' interpersonal skills and attitudes towards group projects (Burdett 2003). This section discusses common group problems, many of which are interrelated and often co-occur (Roberts and McInnerney 2007).

Social loafing. Social loafing occurs when a member consistently contributes less than others. Individuals may be less productive in a group than working alone (Ringelmann 1913; Ingham et al. 1974). This phenomenon decreases the group's ability to perform to its potential, resulting in unwarranted marks and morale damage (Karau and Williams 1993; Latané et al. 1979). Individual monitoring and assessment may reduce this (Williams et al. 1981).

Different attitudes and expectations. The attitude can determine a student's role within the project, the degree of participation and motivation, the grade that is aimed for and much more. Attitudes can be influenced by experience and age (Barfield 2003), and they are closely tied in with the expectations about group work. A group with many clashing attitudes or different expectations about the group work and the final product can present fundamental conflicts (Mackie and Goethals 1987).

Communication problems. Communication is essential while working collaboratively (MacMillan et al. 2004). Teams with poor and ineffective communication among their members perform considerably less compared to other teams (Cervone 2014). Poor communication can be caused by a member's lack of interpersonal skills or inexperience, or cultural differences (Lolli 2013; Liu et al. 2010).

Diversity. Schoenecker et al. (1997) refer to diversity in classrooms as differences that induce different behaviours, attitudes, norms and/or communication patterns. While some types of diversity are known to be beneficial for group work (Dahlin et al. 2005), high diversity in authority or social power (defined as *disparity* by Harrison and Klein (2007)) may result in group competition, resentment following member input, and withdrawal from the group (Tost et al. 2013; Garandeau et al. 2014).

1.2. Group assessment attributes

Group problems can be challenging to measure by themselves. In this section, we introduce five attributes that dissect group work into measurable variables. We will base our assessment attributes on the ones Phielix et al. (2011) used to raise awareness of bias, namely: Quality of Contribution, Productivity, Reliability, Cooperation, Friendliness, and Influence. *Quality of Contribution* depends on factors such as the level of writing or the depth and complexity of the contribution. *Productivity* refers to the quantity of work an individual contributes within a certain time frame. Social loafing and group disparity are known to result in unequal contributions (Karau and Williams 1993; Harrison and Klein 2007). *Friendliness* can be defined as the willingness to help other members, being considerate and friendly, respectful, and inspiring and trusting others (Wubbels et al. 1985). *Cooperation* is the stance a member takes while participating in a group project (Wubbels et al. 1985). The more a student leans towards cooperation, the better the group cohesiveness, which is known to have an enhancing effect on performance (Craig and Kelly 1999; Huang 2009). *Reliability* is the cognitive and emotional assurance that the interests of fellow members are respected and that the individual is working towards the benefit of the group (Emans et al. 1996). *Influence* reflects the role or attitude that a team member might assume during a project (Bales 1988). This attribute is less suitable to identify group problems given its complexity, and so excluded.

1.3. Assessment methods

This paper focuses on computer-supported peer assessment, where every individual provides feedback about the other members. Peer assessment is suitable for assessing group dynamics in student teams (Martinazzi 1998). It has been linked to improvement in collaboration, communication, and cooperation among students (Issa 2012). Furthermore, computer-supported assessments have been found to make students more comfortable when answering class-related questions (Sharma et al. 2005). Several computer-based peer assessment systems have been developed. For instance, *SPARK* allows students to rate each other on multiple criteria in a typical assessment form, which can be adapted to the course needs (Freeman and McKenzie 2002). *WebPA* works similarly, but the result is run through an algorithm to return the final score of a student (Loddington et al. 2009). Lastly, in the *LearnEval* tool by Badea and Popescu (2019), the performance of a student is derived from the scores received from the teachers, peers, and general metrics of the course (e.g. the number of reviews submitted). These systems are not strongly linked to group assessment attributes, and leave this up to teachers.

Anonymity in peer assessment may reduce the degree of social desirability bias, leading to more accurate and valid peer reviews (Wildman 1977), but can reduce positive effects of peer assessment, such as increased contributions (Bamberger et al. 2005). A *self-measure* may produce a competitive environment, where students feel that negatively assessing their peers' performance compared to their own is self-beneficial, impacting group cohesion and dynamics (Bamberger et al. 2005).

2. METHODS

Three qualitative studies were conducted. Exploratory interviews were conducted to investigate how the problems mentioned in the literature apply to Dutch students. Then, focus groups were used to gather data about the effectiveness of two possible survey visualizations. Lastly, the resulting surveys were tested in a field study.

2.1. Exploratory Interviews

Cultural and contextual differences may influence the type of problems and the way in which group members are assessed. The goal of the interviews was to ensure a good understanding of the research context, checking how the problems mentioned in the literature apply to Dutch students.

2.1.1. Interview Design

Six semi-structured interviews (about 20 minutes each) took place using video-conferencing software; one was held face-to-face. Participants received a drink or snack as compensation. Interviews were audio recorded, and later transcribed. After providing informed consent, participants were explained the scope of the interview and procedure. The rest of the interviews explored the following topics: (a) Problem experiences, (b) Assumed causes, (c) Problem resolution methods, (d) Group assessment experiences, (e) Suggested improvements for peer assessment and group projects, (f) Anonymity, (g) Self-measurement.

2.1.2. Participants

Participants were gathered through convenience sampling. They had to be between 18 and 30 years old, and have been a student in the Netherlands no longer than two years ago. Seven (ex-)students (4 male, 3 female; age M=23.3, SD=1.1) participated.

2.1.3. Results and Discussion

Problems experienced. All mentioned cases of social loafing (12 instances) and poor communication (10 instances), affecting both the quality and quantity of delivered work. Diversity and differences in attitudes were mentioned less frequently (by 4 and 5 participants respectively, for 8 and 7 instances). These may be more difficult to observe from the point of view of a student, who would notice instead their consequences on members' behaviour, as signalled by ineffective communication, tension, and conflicts among members. As mentioned in Section 1.1, many of these problems are interrelated (Roberts and McInnerney 2007). This was reflected by the participants' experiences, who mentioned that social loafing and poor communication often led the other motivated member(s) to do most work. In addition to the problems already mentioned in the literature, the lack of involvement of teaching staff was mentioned by 3 participants (4 instances). This suggests an increased frequency of peer assessments may be needed to encourage involvement of teachers, who would have an additional opportunity to reach out.

Assumed causes of problems. Most were unaware of the reason behind the problematic behaviour of others, hinting toward ineffective communication. One mentioned a general lack of motivation to work on the project. Another blamed it on controlling and dominant behaviour by a group member. In general, frustrating actions of members were often mentioned to convey negative attitudes towards group work. Only two reflected on their own role within the problem, signalling a potential lack of self-reflection.

Problem resolution methods. The main resolution method was involving the teacher. Other strategies included solving problems internally, avoiding conflicts and/or covering for problematic group members. When attempts to solve problems internally were unsuccessful, the teaching staff was involved. When the teachers' involvement was unhelpful, groups often opted for avoiding further conflict, or some members would cover for others' problematic behaviour by increasing their own workload.

Group assessment experiences. The most common assessment methods experienced were logbooks and peer assessments on personal performance. Logbooks were documents that kept track of each member's tasks and the amount of time worked on their respective tasks. Participants were neutral towards the effectiveness of logbooks. The peer assessments were mostly surveys with predetermined categories, where each member received grades for every category. Participants were more positive towards peer assessments, but mentioned issues in implementing these. A first issue was too little space to properly motivate one's assessment. A second problem was timing: peer assessments were often at the end of the course. Participants considered such a survey useful only for ”defending your own grade”. Less used methods were (video) conversation, peer review forms, or giving fellow members a grade for the overall quality of their contribution.

Suggested improvements. Improvements regarding group projects focused mainly on increasing teacher involvement, especially regarding conflict resolution. Additionally, many suggested that students should be allowed to choose their own group members. Suggestions for improving peer assessments included: more frequent measuring moments using forms, getting feedback as a result of submitting feedback forms, and the ability to freely motivate answers instead of being tied to a 5-point scale.

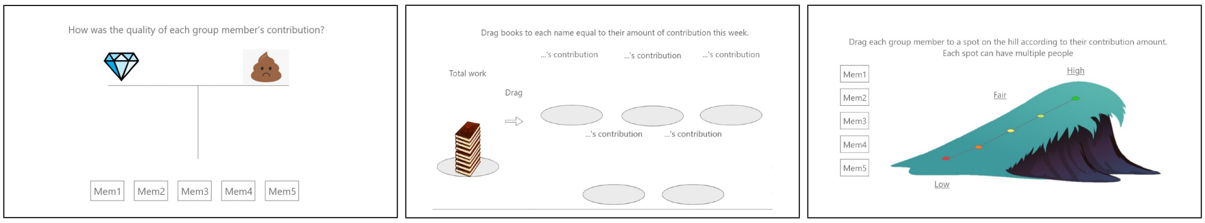
Anonymity. Four participants were pro anonymity, one against, two neutral. In favour of anonymity, it was mentioned that it is easier to be honest when anonymous, and may avoid uncomfortable moments among group members. Noticeably all in favour of anonymity experienced severe problems in project groups. The argument against anonymity was that it may worsen the group climate. Anonymity is discussed further in the focus groups below.

Self-measurement. Four participants were against the inclusion of self-measurement in peer assessment, as it could create a competitive environment - in line with the study of Bamberger et al. (2005). Two were in favour, as it encourages self-reflection. This was also recognized by three participants against self-measurement. On balance, it was decided not to include self-measurement in the peer assessment.

In conclusion, peer assessment seems best to assess group attributes, since most were already familiar with it. Based on improvement suggestions and reported negative experiences, peer assessment should be submitted with a higher frequency than only at the end of the project. After each submission, data should be analyzed and shared with teachers, who can get involved in problem resolution when needed. Also, students should be able to explain their ratings.

2.2. Focus Groups

Focus groups (FGs) were conducted to study which formats, visualizations, and tasks are best received by the target group of students. Besides consistent submission of assessments, it is important to study what contributes to the honesty of submission and what formats are generally liked.


How was the quality of each group member's contribution?
Mem1
Mem2
Mem3
Mem4
Mem5
Drag books to each name equal to their amount of contribution this week.
...'s contribution
...'s contribution
...'s contribution
Total work
Drag
...'s contribution
...'s contribution
Drag each group member to a spot on the hill according to their contribution amount.
Each spot can have multiple people
Mem1
Mem2
Mem3
Mem4
Mem5
High
Fair
Low

***(a)** Binary assessment*

***(b)** Point distribution*

***(c)** Likert scale with visual elements*

***Figure 1:** Advanced visualizations examples with three formats: binary assessment (Fig. 1a), point distribution (Fig. 1b) and 5-point Likert scale with visual elements (Fig. 1c).*

2.2.1. Visualizations Design

Two types of visualization of group assessment questions were created.

Advanced visualizations. The advanced visualizations focused on keeping the respondent engaged and motivated by including visually pleasing elements (e.g. colour) and increased interactivity (e.g. drag and drop). They were designed with three formats in mind: binary assessment, point distribution, and a creative form of the 5-point Likert scale. Binary assessment required one of two answer options (e.g. 'Yes' or 'No'); an example can be seen in Fig. 1a. In point distribution an X-number of points is distributed amongst, or assigned to, each group member (Fig. 1b). Lastly, the standard 5-point Likert scale is enriched with visually pleasing or engaging elements, as can be seen in Fig. 1c. For each of the three formats, one visualization was created per attribute, resulting in 15 unique visualizations.

Basic visualizations. The basic visualizations were designed with the goal of producing an easy and time-efficient assessment, with simple elements that require a single click. Six basic visualizations were created, each with different factors, such as the addition of images, assessing one student per screen versus multiple students per screen, and the type of input required. Fig. 2 shows three examples.

2.2.2. Participants

Participants for the two FGs were gathered using a combination of convenience and snowball sampling. Inclusion criteria were: must be between 18 and 30 years old, be or have been a student in the Netherlands no longer than two years ago, and must have experience in group projects in some form of higher education. There were two FGs of four participants (FG1 and FG2), so eight participants in total. Participants were between 23 and 24 years old (M=23.88, SD=0.74). Four were male, four female.

2.2.3. Procedure

FGs were held in a combination of an online visual collaboration platform called Miro (www.miro.com) and a video conferencing tool. Using Miro, each participant could interact with elements placed in the virtual whiteboard space. During the sessions, the researcher oversaw the FG, explained the tasks, and took notes. First, participants joined the video call, were briefly introduced to the research and what would happen in the FG. They provided informed consent and were introduced to Miro functionalities.

Part 1: Basic visualizations. Participants ranked the six basic visualizations from 1 (best) to 6 (worst) on what would (a) motivate them to keep submitting feedback, (b) motivate them to answer truthfully, and (c) preferred overall. This ranking on a-c was done per attribute resulting in 15 rankings. When each participant had completed their individual rankings, the FG discussed these and produced a collaborative ranking. A 10 minute break was given.

Part 2: Advanced visualizations. Because of time constraints, FGs ranked the three advanced visualizations per attribute per topic, instead of integrating them into the basic visualization rankings. After the ranking was completed, questions were asked about the motivations behind the ranking.

Part 3: Group discussion. First, FGs discussed which question order would be best and dragged the five attributes into the preferred order to present them in a peer assessment form. Second, FGs discussed anonymity towards fellow group members. Participants put pros and cons on sticky notes and discussed what would be more beneficial: submit the assessment anonymously or not.

2.2.4. Results and Discussion

Basic visualizations. Each visualization was given points from 0 (last place) to 5 (first place). The individual participant scores were combined into a total per FG per topic (motivation to submit, motivation to be honest, and preference). Given that participants of both FGs considered all topics equally important, the total scores per visualization were summed up across topics, resulting in one final ranking per visualization per each attribute. Both FGs agreed on the visualization to use for Reliability (Fig. 2b) and Quality of Cooperation and Contribution (Fig. 2a), but were divided for Productivity and Friendliness. In case of disagreement, the scores of

FG1 and FG2 were averaged, and the visualization with the highest average was used. This resulted in choosing the 5-point Likert scale with stars for Quality of Cooperation, Quality of Contribution, and Productivity; the 5-Point Likert scale with dots for Reliability, and the 5-point slider with emoticons for Friendliness (Fig. 2c). Both FGs' high ranking on the 5-point Likert scale with stars is likely caused on the one hand by it being a simple and quick format, and on the other hand by the stars giving a more positive twist to the assessment (as commented by participants). Another highly ranked visualization was the one in Fig. 2b, which received high scores on every attribute but Friendliness, which may suggest lesser suitability for personality-related questions.

Advanced visualizations. Ranking was done similarly: for each attribute, the three corresponding visualizations received a score from 0 to 2. Both FGs preferred the 5-point scale variant (Fig. 1c) for most attributes. This suggests that participants preferred a middle ground between speed and detail. For Quality of Contribution FG2 preferred point distribution and FG1 the 5-point Likert scale. We decided to use point distribution for Quality of Contribution and the 5-point Likert scale for other attributes.

Attribute order. The order was fairly similar for both FGs, but they took a different approach. While FG1 ordered the attributes according to how personal they felt, FG2 took the importance of an attribute into account. The only difference was between Quality of Contribution and Productivity, which were ordered third and second by FG1, and second and third by FG2, respectively. Both groups admitted that these attributes strongly relate. Combining the results of the two FGs suggests that less personal attributes (e.g. Quality of Cooperation) are more important than highly personal attributes (e.g. Friendliness). This will be further investigated in the field research.

Anonymity. FGs were divided, with FG1 against and FG2 in favour. Main arguments in favour of anonymity related to honesty and feeling safe to express your opinion without risking group cohesiveness. The arguments against anonymity were that it could hinder personal growth, discourage communication, that it gives people the option to say whatever they want without repercussions, and that it works only in larger groups. The group against anonymity had experienced most group problems, suggesting students with severe problems feel more comfortable when answers are kept anonymous. This result is in line with the interviews results and the studies of Wildman (1977) and Li (2017). So, we will use anonymity towards fellow group members.

2.3. Design of Peer Assessment Survey

Based upon the results of the interviews and focus groups a peer assessment survey was designed. It was designed to be short and to the point, but to leave space for free text to motivate answers.

Visualizations. The top-rated basic- and advanced visualizations were used to create two versions of the survey: Version A (advanced visualization) and Version B (basic visualization). Each version had one question for each of the five attributes.

Order. The order in which the attributes were assessed was the one suggested by FG2: Quality of Cooperation, Quality of Contribution, Productivity, Reliability, and Friendliness.

Anonymity. The assessment survey was made anonymous towards the participant's fellow group members, while keeping it transparent to the teachers. This was decided to encourage groups with problems to fill out the assessment honestly and frequently. The anonymity would not extend to teachers, as they are expected to monitor the students and potentially intervene.

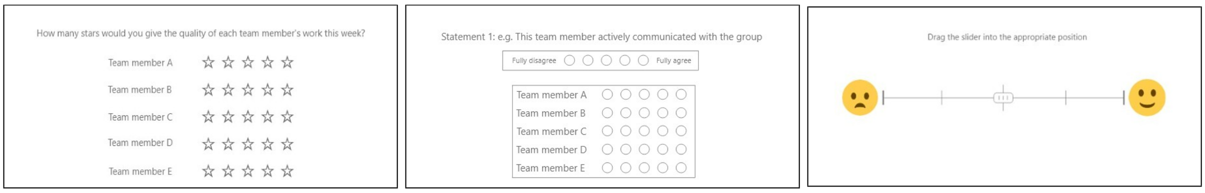
Self-assessment. No self-measurement was included in the survey, to avoid creating a competitive environment and an overestimation of someone's own contribution to the project.

Implementation. Both versions of the peer assessment were created using the Utrecht University's Qualtrics environment. The digital sketches of the visualizations were put into a real survey format by using the Qualtrics toolbox. Unfortunately, it was not possible to create an exact replica of all visualizations. This was primarily an issue for the advanced visualizations, where the drag-and-drop interaction had to be slightly modified, and did not fit entirely on mobile device screens. To avoid this issue, students were asked to fill out the surveys on a computer.

2.4. Design of E-Mate for initial support

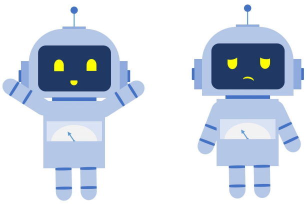
An important part of solving group conflicts consists in decreasing the negative emotions arising from them. A further advantage of software-supported peer assessment is the possibility of delivering a simple form of immediate emotional support, offering a temporary relief from the stress connected to possible group problems. Therefore, right after a survey was filled in, an initial reflection on the results was provided by E-Mate, a virtual agent with robot design. An additional benefit of such a system would be to offer immediate feedback after the survey, given that students mentioned the lack of feedback on peer assessments as potential for improvements.

Feedback. Feedback was designed with two main goals: creating feedback that effectively reflects the magnitude of issues students encountered, and


How many stars would you give the quality of each team member's work this week?
Team member A ☆ ☆ ☆ ☆ ☆
Team member B ☆ ☆ ☆ ☆ ☆
Team member C ☆ ☆ ☆ ☆ ☆
Team member D ☆ ☆ ☆ ☆ ☆
Team member E ☆ ☆ ☆ ☆ ☆
Statement 1: e.g. This team member actively communicated with the group
Fully disagree ○ ○ ○ ○ ○ Fully agree
Team member A ○ ○ ○ ○ ○
Team member B ○ ○ ○ ○ ○
Team member C ○ ○ ○ ○ ○
Team member D ○ ○ ○ ○ ○
Team member E ○ ○ ○ ○ ○
Drag the slider into the appropriate position

***(a)** 5-point Likert scale with stars* ***(b)** 5-point Likert scale with dots* ***(c)** 5-point slider with emoticons*

***Figure 2:** The preferred basic visualizations: Quality of Cooperation, Quality of Contribution, and Productivity were paired with 2a; Reliability with 2b, and Friendliness with 2c.*



(a) ***(b)***

***Figure 3:** E-Mate expressions: happy (a) and sad (b).*

offering a first exploration of emotional support tailored to group problems. A basic form of emotional support was provided based on the verbal strategies of exploration of the problem, empathy and reflection. Exploration of the topic has been reported as an important step in emotional support dialogue systems (Liu et al. 2021), recommended especially early in the conversation to signal interest and concern. Empathic messages were conveyed (e.g. *"I'm sorry to hear that"*). Sympathy for the situation was conveyed via reflection (*"That must be very frustrating right now"*). Sometimes, the E-Mate offered encouragement (*"Keep up the good work!"*).

Mediator. The E-Mate proposed itself as a mediator between students and teachers, promising to contact the teachers when the group scores were too low (*"I'll tell the teachers, as I think you need help."*). This was created to facilitate the feeling of being heard and seen by the teachers, given that the interviews mentioned the issue of low teacher involvement.

Feedback styles. Lastly, two possible feedback styles, reported in Table 1, were analyzed: an Attribute-Centered style, proceeding by analyzing each attribute results, one at a time, and an Person-Centered feedback style, where the results were discussed per group member.

Appearance. The E-Mate was designed to avoid overestimation of its abilities, as a more human-like appearance may lead to expectations of human-like empathic skills, and so in users' disappointment when faced with robotic-like feedback (Smith and Masthoff 2018; Go and Sundar 2019). At the same time, its appearance needed to suggest that some form of basic empathy was possible. Therefore, it exhibited also a simple form of emotional expression, displaying a happy or a sad face and body according to the overall mean scores (see example in Fig.3).

2.5. Field Study

The peer assessment including the E-Mate was evaluated in a field study.

2.5.1. Field Study Design

The field study happened in a course in which students worked in project groups. A between-subject design was used: members of half the groups received peer assessment Version A, the others B. The assessment was done twice[1] in course weeks 4 and 7, followed by a feedback survey at the course end. Assessments were soon after a group deadline to not impede on deadlines and give teachers enough time to possibly intervene before future group work. A criterion was used to identify whether a group was experiencing a problem. In the first survey, this was when one member received a score $\leq$3 on three attributes. In the second survey, this was reduced to $\leq$3 on two attributes, as we found that students generally gave high scores.

2.5.2. Feedback Survey

This survey obtained feedback on the peer assessments. After providing informed consent, participants indicated how many assessments they filled out, and if so the version received, and whether their group experienced problems. Next, they rated the appropriateness of the attributes for measuring group problems. Then, they indicated their agreement level with statements assessing: (a) Suitability of the format for the attributes; (b) How the format motivates (i) to respond truthfully, (ii) finish the survey and (iii) submit future surveys; (c) Appropriateness of the length and the difficulty of the survey; (d) Usefulness for the teachers; (e) Satisfaction with the implementation of anonymity; (f) Need for self-measurement; (g) Opportunity for self-reflection. They commented on what they liked and what could be improved in the survey and group projects. Lastly, they reflected on the E-mate with regard to preferred feedback format (Attribute-Centered or Person-Centered), their feelings about it, and possible improvements. Participants who had submitted no assessments explained why and what would motivate them to submit in future.

[1] Members received the same version in each round

Table 1: *Attribute-Centered and Person-Centered feedback. In this example, someone rated all attributes for all members positively, except for Mario who was rated neutral on Quality of Cooperation and Production, and negatively on Reliability.*

Attribute-Centered feedback	**Person-Centered feedback**
It sounds like things are quite amazing between you guys! I am glad that most of the other members are *cooperating* well with you. It's great that you think everybody is providing *high-quality contributions* to the project. Happy to hear that most other members are *productive*. Glad to hear most of the other members are easy to *rely* on, although I am sorry to hear that Mario is not so reliable. Similarly, I'm so glad to hear you *get along* with everybody. Getting along makes things always better.	It sounds like you guys make quite an amazing team! It's great that things are so nice with most team members. *You guys* seem to cooperate well, to provide good quality contribution and to be very productive together. It's also great that you are getting along and relying on each other: this is of great importance while working together. With most of the work going so well, it's a pity that you are still experiencing some minor issues. I'm sorry to hear that things could use some improvement with *Mario*.
Should any problem arise, let us know, because it's very important that the teachers are aware of any problems in your group. Keep up the good work!	

2.5.3. Procedure

Teachers emailed students about the peer assessment, and that they had 7 days to fill it out. The survey provided general information about what to expect and the option to share data with only the teachers or also the researchers. Then, questions about each attribute were asked; these contained response fields for each fellow group member with a comments field. Finally, the E-mate gave Attribute-Centered feedback on the answers just given. For the second assessment, the explanation of Reliability was extended, as some felt it was not fully clear. Furthermore, the E-Mate feedback changed to Person-Centered, and included a comparison with the scores given in the first round, commenting on potential improvements or deteriorations. After each round, a teacher reviewed the data and potential group problems communicated to other teachers, with advice to reach out to any problematic groups. Finally, the feedback survey was sent to students and teachers were asked by email whether they found the peer assessment useful, would be willing to use it in future, and had any improvement suggestions.

2.5.4. Participants

Participants were students enrolled in a bachelor course. All responses were voluntary. For the feedback survey participants had a chance to win one of two €10,- gift cards. 87 students in total were enrolled in the course. In the first assessment round, 53 responded, of which 34 agreed to share data with the researcher. In the second round 29 responded, of which 20 agreed to share data. 25 completed the feedback survey (plus 7 excluded as incomplete).

2.5.5. Results

Problem identification. Four groups met the problem identification criterion in the first round. In one of these, comments were added, noting problems in communication and work ethics. These problematic cases were communicated to the teachers, who reached out to these groups. In the second round, three met the problem identification criterion, two of which had reported problems in the first round as well. These were similarly signalled to the teachers. So, the peer assessment effectively identified potential problematic groups, which ultimately led to teachers reaching out to them.

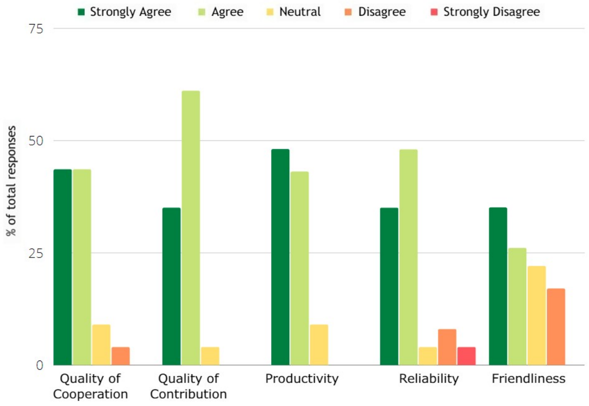


Figure 4: *Results of Attribute Appropriateness.*

Feedback survey. Of the 25 who filled out the survey, 18 had filled out both assessments, 5 only one, 2 none. 8 had received version A, 15 B. 10 reported having experienced a problem in their group. Three experienced unequal contributions, 5 communication issues, and one mentioned attitude problems. Other problems were an unpleasant atmosphere and unclear assignment descriptions. From the two that had not filled out any assessments, one mentioned that there was no incentive to fill it out and the other that the survey would most likely not have any benefit to them, as they did not have any problems in their group. The 5 that had filled out only one survey mostly forgot to fill out the other.

Attributes' appropriateness and usage. Most considered the attributes appropriate (Fig. 4). Opinions were more divided for Reliability and Friendliness. No comments elaborated on Reliability. In line with the FGs, Friendliness was considered by some as highly subjective and of lesser importance. Table 2 shows the proportion who used an attribute to distinguish between group members. All attributes are used, but Friendliness less.

Question format. Figure 5 compares the two versions. The sample size did not allow for meaningful statistical analyses. To ease comparison, statement responses were converted into numbers from -2 (strongly disagree) to 2 (strongly agree). The median of both versions and overall are shown in Table 3. Version B's participants thought that the question format was *suitable* for the attributes, whilst A's were more divided. Most participants of both versions agreed that the format motivated them to answer *truthfully*, though many answers were neutral for Version B. On whether the format motivated them to *finish* the survey, Version A's participants were mostly neutral and B's more positive. Overall, opinions were divided on whether the format motivated submitting *future* surveys. Most strongly agreed that the survey's *length* and *difficulty* were appropriate. Overall, Version B received slightly more positive answers than A. This may be due to the reduced usability of the drag-and-drop function in Version A when implemented in Qualtrics.

Perceived usefulness by students. Fig. 6 shows statements for which no distinction was needed between versions. Students agreed on the *usefulness* of the survey, that it created an opportunity for self-*reflection*, and that the implementation of *anonymity* was satisfactory. Most disagreed that it led to a group internally solving its *problems*, but this was to be expected, as the number of problems experienced was low and no feedback was given yet to group members on what others thought or how to solve problems. Most desired self-measurement (*Self-M*), perhaps as they wished to defend their self-interest.

Liked aspects. Participants appreciated the easy-to-understand and quick-to-complete format, regardless of the survey version. Three mentioned positive feelings resulting from a teacher reaching out to them. Other positive aspects were the anonymity and the positive influence of Version A on motivation.

Improvements. Some participants suggested reducing the survey length, as the questions were time-consuming, despite the results on length appropriateness (Fig. 5). Other suggestions were to explain the survey relevance more before distributing it, give tips on how to improve collaboration, remove the Friendliness question, and change the drag-and-drop interaction in Version A. Regarding improvements to group projects, some suggested making peer assessments mandatory, and interacting more with the teams during the project.

Feedback teachers. Four teachers responded. All thought the peer assessment was helpful and would like to use this assessment form in future courses. One mentioned that the assessment was useful to some extent, but their teams did not experience many problems or had not filled out any assessment. Suggested improvements included changes to question phrasings and the addition of a question asking if the student wants the teacher to intervene. Another suggestion was to include a self-evaluation and a general question along the lines of *"How do you think your group is doing?"*, to help disambiguate between cases who need help and students who tend to use lower scores.

E-Mate feedback style. 16 out of 21 respondents preferred the Person-Centered to the Attribute-Centered feedback, describing it as more honest, clearer, more natural, and more adapt to describe a problem with a specific group member. Negative aspects of the Person-Centered feedback were also mentioned: it was perceived by some as too negative and robotic. The Attribute-Centered feedback was similarly reported as too automated, too mean, and too fake, although some preferred it and perceived it as friendlier, more positive and human-like.

E-Mate effect. Participants were equally divided between positive, neutral and negative. The positive opinion was that it constituted a nice personal touch; neutral that it was okay, but did not contribute much. Negative opinions mostly referred to it not being useful, as it resembled a summary too much. One participant mentioned that it was accurate, suggesting that the E-Mate effectively captured the problems experienced by their group. Suggested improvements included changes in phrasing, such as making it more of a story, more human and personal, and adding tips on how to solve problems.

***Table 2:** Percentage of participants discerning between group members using each group work attribute.*

Attribute	Round 1	Round 2
Reliability	36.3%	45%
Quality of Cooperation	51.5%	45%
Quality of Contribution	51.5%	65%
Productivity	36.3%	65%
Friendliness	27.3%	35%

***Table 3:** Median scores of the feedback statements.*

Statement	Version A	Version B	Overall
Suitability	0.5	1	1
Truth	1	1	1
Finish	0	1	0.5
Future	0	0	0
Length	1.5	2	2
Difficulty	2	2	2
Usefulness	1	1	1
Problem	-0.5	-1	-1
Anonymity	1	2	2
Self-M	1	1	1
Reflection	1	1	1

3. CONCLUSION

This research investigated the detection of group work problems via computer-supported assessments. In interviews and focus groups the most

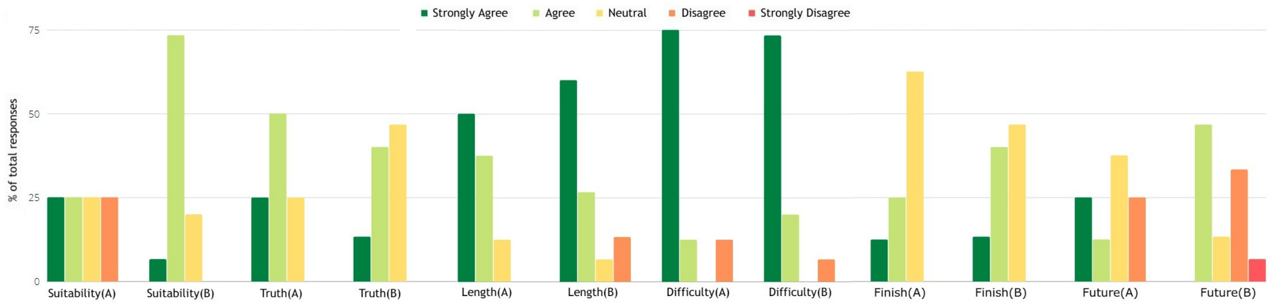


Figure 5: *Comparison between Versions A and B.*

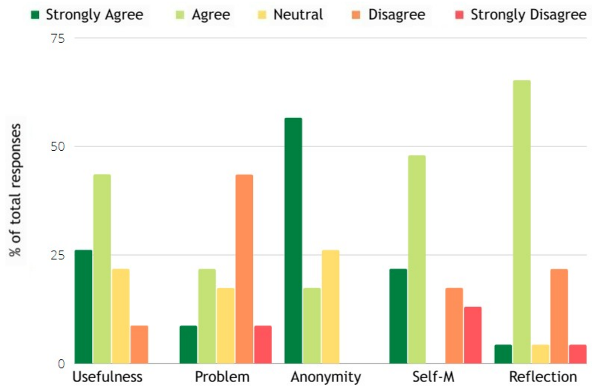


Figure 6: *Results of Usefulness, Problem Resolution, Anonymity, Self-Measurement, Reflection.*

common group issues were investigated and which survey features are best received by students. A peer assessment survey was chosen as most suitable to implement the improvements suggested in both interviews and focus groups. Two surveys with visually diverse elements were created. No significant differences in promoting honesty, motivation and general liking were found between the two.

The research extended the work of Phielix et al. (2011) by investigating the suitability of their attributes (aimed at increasing awareness of judgement biases) for detection of group issues, studying different visualisations, and exploring the use of E-Mate to give initial feedback. The attributes were considered suitable by students and used to differentiate between group members, though some regarded Friendliness as subjective and less important. Overall, the survey was well received by both students and teachers, who deemed it a useful method to assess group work. The peer assessment resulted in the identification of three groups experiencing problems, of which at least two required teacher intervention.

The field study only considered one course. Similar studies with more courses, and larger sample sizes are needed to generalize the results. We will conduct more research to improve the E-mate and measure its effect on students' stress. Additionally, we will investigate how to visualize the data for teachers and groups, to better support problem resolution.

REFERENCES

Adeniran, A. (2020), Investigating Real-Time Assessment and Support for Online Collaborative Learning, PhD thesis, University of Aberdeen.

Badea, G. and Popescu, E. (2019), Instructor support module in a web-based peer assessment platform, *in* 'Proc. System Theory, Control and Computing', IEEE, pp. 691–696.

Bales, R. (1988), *Overview of the SYMLOG system: Measuring and changing behavior in groups*, SYMLOG Consulting Group Woodland Hills, CA.

Bamberger, P. A., Erev, I., Kimmel, M. and Oref-Chen, T. (2005), 'Peer assessment, individual performance, and contribution to group processes: The impact of rater anonymity', *Group & Organization Management* **30**(4), 344–377.

Barfield, R. L. (2003), 'Students' perceptions of and satisfaction with group grades and the group experience in the college classroom', *Assessment & Evaluation in Higher Education* **28**(4), 355–370.

Burdett, J. (2003), 'Making groups work: University students' perceptions', *International Education Journal* **4**(3), 177–191.

Cervone, H. F. (2014), 'Effective communication for project success', *OCLC Systems and Services: International digital library perspectives* .

Colbeck, C. L., Campbell, S. E. and Bjorklund, S. A. (2000), 'Grouping in the dark: What college students learn from group projects', *The Journal of Higher Education* **71**(1), 60–83.

Craig, T. Y. and Kelly, J. R. (1999), 'Group cohesiveness and creative performance.', *Group dynamics: Theory, research, and practice* **3**(4), 243.

Dahlin, K. B., Weingart, L. R. and Hinds, P. J. (2005), 'Team diversity and information use', *Academy of management journal* **48**(6), 1107–1123.

Emans, B., Koopman, P., Rutte, C. and Steensma, H. (1996), 'Teams in organisaties', *Gedrag en Organisatie* **6**, 309–327.

Freeman, M. and McKenzie, J. (2002), 'Spark, a confidential web–based template for self and peer assessment of student teamwork: benefits of evaluating across different subjects', *British journal of educational technology* **33**(5), 551–569.

Garandeau, C. F., Lee, I. A. and Salmivalli, C. (2014), 'Inequality matters: Classroom status hierarchy and adolescents' bullying', *Journal of youth and adolescence* **43**(7), 1123–1133.

Go, E. and Sundar, S. S. (2019), 'Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions', *Computers in Human Behavior* **97**, 304–316.

Harrison, D. A. and Klein, K. J. (2007), 'What's the difference? diversity constructs as separation, variety, or disparity in organizations', *Academy of management review* **32**(4), 1199–1228.

Huang, C.-C. (2009), 'Knowledge sharing and group cohesiveness on performance: An empirical study of technology r&d teams in taiwan', *Technovation* **29**(11), 786–797.

Ingham, A. G., Levinger, G., Graves, J. and Peckham, V. (1974), 'The ringelmann effect: Studies of group size and group performance', *Journal of experimental social psychology* **10**(4), 371–384.

Issa, T. (2012), 'Promoting learning skills through teamwork assessment and self/peer evaluation in higher education.', *International Association for Development of the Information Society* .

Karau, S. J. and Williams, K. D. (1993), 'Social loafing: A meta-analytic review and theoretical integration.', *Journal of personality and social psychology* **65**(4), 681.

Latané, B., Williams, K. and Harkins, S. (1979), 'Many hands make light the work: The causes and consequences of social loafing.', *Journal of personality and social psychology* **37**(6), 822.

Li, L. (2017), 'The role of anonymity in peer assessment', *Assessment & Evaluation in Higher Education* **42**(4), 645–656.

Liu, L. A., Chua, C. H. and Stahl, G. K. (2010), 'Quality of communication experience: Definition, measurement, and implications for intercultural negotiations.', *J. of Applied Psychology* **95**(3), 469.

Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., Jiang, Y. and Huang, M. (2021), 'Towards emotional support dialog systems', *arXiv preprint arXiv:2106.01144* .

Loddington, S., Pond, K., Wilkinson, N. and Willmot, P. (2009), 'A case study of the development of webpa: An online peer-moderated marking tool', *British J. of Educational Tech.* **40**(2), 329–341.

Lolli, J. C. (2013), 'Interpersonal communication skills and the young hospitality leader: Are they prepared?', *Int. J. of hospitality management* **32**, 295–298.

Mackie, D. M. and Goethals, G. R. (1987), Individual and group goals., *in* C. Hendrix, ed., 'Group processes', Sage Publications, Inc.

MacMillan, J., Entin, E. E. and Serfaty, D. (2004), Communication overhead: The hidden cost of team cognition., *in* E. Salas and S. M. Fiore, eds, 'Team cognition: Understanding the factor that drive process and performance', APA, pp. 61–82.

Martinazzi, R. (1998), Design and development of a peer evaluation instrument for" student learning teams", *in* 'Proc. of Frontiers in Education Conference.', Vol. 2, IEEE, pp. 784–789.

Murray, J.-A. and Boyd, S. (2015), 'A preliminary evaluation of using webpa for online peer assessment of collaborative performance by groups of online distance learners.', *Int. Journal of E-Learning & Distance Education* **30**(2), n2.

Phielix, C., Prins, F. J., Kirschner, P. A., Erkens, G. and Jaspers, J. (2011), 'Group awareness of social and cognitive performance in a cscl environment: Effects of a peer feedback and reflection tool', *Computers in human behavior* **27**(3), 1087–1102.

Ringelmann, M. (1913), *Recherches sur les moteurs animés. Travail de l'homme.*, Annales de l'Institut national agronomique.

Roberts, T. S. and McInnerney, J. M. (2007), 'Seven problems of online group learning (and their solutions)', *Journal of Educational Technology & Society* **10**(4), 257–268.

Schoenecker, T. S., Martell, K. D. and Michlitsch, J. F. (1997), 'Diversity, performance, and satisfaction in student group projects: An empirical study', *Research in Higher Education* **38**(4), 479–495.

Sharma, M. D., Khachan, J., Chan, B. and O'Byrne, J. (2005), 'An investigation of the effectiveness of electronic classroom communication systems in large lecture classes', *Australasian Journal of Educational Technology* **21**(2).

Smith, K. A. and Masthoff, J. (2018), Can a Virtual Agent provide good Emotional Support?, *in* 'Proc. British Human Computer Interaction', pp. 1–10.

Tost, L. P., Gino, F. and Larrick, R. P. (2013), 'When power makes others speechless: The negative impact of leader power on team performance', *Academy of Management J.* **56**(5), 1465–1486.

Wildman, R. C. (1977), 'Effects of anonymity and social setting on survey responses', *Public Opinion Quarterly* **41**(1), 74–79.

Williams, K., Harkins, S. G. and Latané, B. (1981), 'Identifiability as a deterrant to social loafing: Two cheering experiments.', *Journal of Personality and Social Psychology* **40**(2), 303.

Wubbels, T. et al. (1985), Discipline problems of beginning teachers, interactional teacher behaviour mapped out., *in* 'Resources in Education', ERIC.