


How well do collaboration quality estimation models generalize across authentic school contexts?

Pankaj Chejara | Reet Kasepalu | Luis P. Prieto |
María Jesús Rodríguez-Triana  | Adolfo Ruiz Calleja |
Bertrand Schneider

Pankaj Chejara, Tallinn University, Tallinn, Estonia

Correspondence

Pankaj Chejara, Tallinn University, Tallinn, Estonia.

Email: pankajch@tlu.ee

Funding information

Estonian Research Council's Personal Research Grant, Grant/Award Number: PRG1634; European Union's "NextGenerationEU/PRTR"; Spanish Ministry for Science and Innovation, Grant/Award Number: RYC2021-032273-I; Spanish Ministry for Science and Innovation, Grant/Award Number: PID2020-112584RB-C32

Abstract

Multimodal learning analytics (MMLA) research has made significant progress in modelling collaboration quality for the purpose of understanding collaboration behaviour and building automated collaboration estimation models. Deploying these automated models in authentic classroom scenarios, however, remains a challenge. This paper presents findings from an evaluation of collaboration quality estimation models. We collected audio, video and log data from two different Estonian schools. These data were used in different combinations to build collaboration estimation models and then assessed across different subjects, different types of activities (collaborative-writing, group-discussion) and different schools. Our results suggest that the automated collaboration model can generalize to the context of different schools but with a 25% degradation in balanced accuracy (from 82% to 57%). Moreover, the results also indicate that multimodality brings more performance improvement in the case of group-discussion-based activities than collaborative-writing-based activities. Further, our results suggest that the video data could be an alternative for understanding collaboration in authentic settings where higher-quality audio data cannot be collected due to contextual factors. The findings have implications for building automated collaboration estimation systems to assist teachers with monitoring their collaborative classrooms.

KEYWORDS

collaboration quality, computer-supported collaborative learning, generalizability, machine learning, multimodal learning analytics

Practitioners notes

What is already known about this topic

- Multimodal learning analytics researchers have established several features as potential indicators for collaboration quality, e.g., speaking time or joint visual attention.
- The current state of the art has shown the feasibility of building automated collaboration quality models.
- Recent research has provided preliminary evidence of the generalizability of developed automated models across contexts different in terms of given task and subject.

What does this paper add

- This paper offers collaboration indicators for different types of collaborative learning activities in *authentic* classroom settings.
- The paper includes a systematic investigation into collaboration quality automated model's generalizability across different tasks, types of tasks and schools.
- This paper also offers a comparison between different modalities' potential to estimate collaboration quality in *authentic* settings.

Implications for practice

- The findings inform the development of automated collaboration monitoring systems for authentic classroom settings.
- This paper provides evidence on across-school generalizability capabilities of collaboration quality estimation models.

INTRODUCTION

"We should now move on to preparing the meal plan", a student in the group suggested during a face-to-face collaborative learning problem where students were asked to plan a class trip (choosing the venue, preparing a travel itinerary, meal options, etc). Other students unanimously agreed with the suggestion, given the 10 minutes remaining from the allotted time of 30 minutes. The students in the group then started suggesting different types of meal options including vegetarian, non-vegetarian and vegan. During this process, an argument arose on whether to keep all three options or not. This created tensions among the group members, triggering negative emotions (e.g., frustration). The situation remained unresolved and the allotted duration ended before completing the task.

This example shows that collaboration is a multifaceted construct, having multiple underlying dimensions (e.g., time management, cooperation, argumentation; Rummel et al., 2011). Various group processes emerge during collaboration (Webb, 2009), which can be beneficial or detrimental to learning. Beneficial processes often include detailed explanations exchanged between peers (Gillies, 2019), help-seeking and help-giving (Webb, 2009) and asking questions to clarify misconceptions. Detrimental processes include a lack of coordination among peers (Barron, 2003), conflicts, social loafing, lack of support from peers, as

well as free-rider, status differential and sucker effects (Salomon & Globerson, 1989). These processes are also likely to trigger participants' emotions in positive and negative ways (Hayashi, 2019). For example, conflicts are likely to result in anger which could negatively impact collaboration if not addressed by the group. These socio-emotional processes can have a negative impact on cognitive processes (e.g., argumentation, knowledge construction) (Huang & Lajoie, 2023). Research has also shown that students' mere participation in collaborative learning activities does not necessarily support learning (Webb, 2009) or result in a successful collaboration (Johnson & Johnson, 1992). It requires self-regulation, socially shared regulation (Hadwin & Oshige, 2011) and external help (King, 2008). Moreover, the help offered needs to be timely and students should have opportunities to use it for the given problem (Webb, 2008).

Research has shown the importance of the teacher in promoting students' active participation, engagement and discussion in collaborative learning groups (Asterhan et al., 2012). To identify and scaffold groups in need, teachers need to be aware of every group's activity during collaborative learning activities (Martinez-Maldonado et al., 2015). However, the complexity of the collaboration process and the multimodal nature of the interaction among students complicate the situation for teachers. Given that some parts of students' cognitive processes are externalized in the form of dialogues, writing and arguments (Stahl, 2006), capturing these artefacts provides opportunities for gaining insight into collaborative processes (Martinez-Maldonado et al., 2019). Multiple streams of data or artefacts of different forms (e.g., written text, spoken text) can potentially facilitate a holistic understanding of collaboration. Multimodal learning analytics (MMLA) has recently emerged as a potential approach towards gaining insight into collaboration processes (Di Mitri et al., 2018; Ochoa, 2017). Though the field is still far from achieving the same level of sensitivity and adaptivity of expert observation, researchers have shown the feasibility of building MMLA-enabled automated detection of high-level collaboration aspects, e.g., collaboration quality, argumentation (Martinez-Maldonado et al., 2015; Pugh et al., 2022).

There is a growing body of research on modelling collaboration, providing supporting evidence on the use of machine learning for estimating collaboration quality (Liu et al., 2021; Reilly & Schneider, 2019; Viswanathan & Vanlehn, 2018). However, the narrow context of those research efforts (each working within one type of learning activity or a single classroom) has offered only a limited understanding of the generalizability aspects of developed models. Besides, given that the majority of these research studies have been conducted in laboratory settings (Chua et al., 2019; Schneider et al., 2022), the knowledge of how well do these developed models perform in *authentic* classroom settings is still missing from current MMLA research. Consequently, there are research gaps on the applicability of automated collaboration models to authentic classroom settings and the generalizability of the developed models in those settings.

To address the aforementioned gaps, this paper sets up the following research questions. **RQ1:** What is the relationship between multimodal data (audio, video, log) features and collaboration quality (and its dimensions) in *authentic* classroom settings? **RQ2:** Whether and to what extent do automated collaboration quality models generalize to different contexts varying on given tasks, types of collaborative learning activities, and schools? **RQ3:** What combination of multimodal data enables the development of a more generalizable collaboration estimation model? By addressing these research questions, this paper takes a step further towards implementing models of collaboration in authentic classroom settings, since little is known about the generalizability of models trained in one setting and applied to other settings.

This paper is structured as follows: Second section summarizes current state-of-the-art research in MMLA on building automated models for collaboration. Third section presents details on our study setup comprising of the study context, data collection tool, dataset

collected, feature extraction and annotation process, model development and evaluation. We report our results in fourth section and discuss their implications for the MMLA community in fifth section. Sixth section, finally, concludes the paper and offers potential future directions for research in this area.

RELATED WORK

Previous research work from learning analytics has provided evidence of the potential of students' log features towards detecting a group's task performance, team's effect and level of collaboration (Goodman et al., 2005; Hernández-García et al., 2018; Yoo & Kim, 2014). For example, Hernández-García et al., 2018 in their study found the distribution of team members' contribution as one of the key indicators for effective team behaviour. Goodman et al., 2005 illustrated the potential of dialogue features (e.g., length of utterance) in detecting collaboration behaviour using neural networks. Furthermore, temporality has also been explored in modelling collaboration. Chounta and Avouris (2012) analysed time-series log data using different time units and identified a 1-minute duration as better for predicting collaboration quality. These works have been further complemented with the use of multimodal data, for example by capturing group interactions from physical space in addition to digital space.

The past decade has witnessed a surge in the use of MMLA for understanding and modelling collaboration behaviour (Chejara et al., 2021; Praharaj et al., 2021; Schneider et al., 2022). Recent research work has even shown that having multimodal data brings performance improvements over mono-modal data for collaboration modelling (Olsen et al., 2020). MMLA researchers have utilized a variety of data sources (such as audio, video, eye gaze and skin-conductance) for modelling collaboration (Pugh et al., 2022; Reilly & Schneider, 2019). From the collected data, features ranging from speaking time, turn-taking and joint-visual attention to facial action units and emotions have been extracted for understanding and modelling collaboration (Cai et al., 2020; Chejara, Prieto, Rodríguez-Triana, Ruiz-Calleja, Kasepalu, et al., 2023; Martínez-Maldonado et al., 2013; Reilly & Schneider, 2019).

From the current state of the art, two main groups of approaches emerged to understand and build automated models for collaboration: statistical methods (Huang et al., 2019) and machine learning (including deep learning) (Spikol et al., 2018; Viswanathan & Vanlehn, 2018). In the first group, researchers mainly used correlation analysis (Huang et al., 2019; Reilly & Schneider, 2019), whereas in the second, researchers employed a variety of algorithms from machine learning, i.e., Decision Tree, Naive Bayes, Support Vector Machine, AdaBoost and Random Forest (Chejara et al., 2021; Martínez-Maldonado et al., 2011; Reilly & Schneider, 2019). The estimation models developed using these algorithms have been found to achieve moderate (75% accuracy) to high performance (84% accuracy) for classifying different levels of collaboration quality (Ponce-Lopez et al., 2013; Reilly & Schneider, 2019). Random Forest algorithm often emerged as the highest-performing collaboration modelling algorithm in MMLA (Ponce-Lopez et al., 2013; Viswanathan & Vanlehn, 2018).

While the majority of the aforementioned research works evaluated their collaboration models with datasets collected from a single context (e.g., a particular activity in laboratory settings), a recent study extended this work by investigating the potential of collaboration automated models for different task contexts in laboratory settings (Pugh et al., 2022). However, to the best of our knowledge, there has not been any cross-evaluation of supervised machine learning models of collaboration across educational settings that differ in terms of their student body, tasks, type of tasks, schools, etc.

From the aforementioned state-of-the-art, we identify three research gaps on the modeling of collaboration using MMLA. First, there is a lack of knowledge of the relationship between multimodal data and collaboration quality (and its dimensions) in *authentic* classroom settings. Second, the current state-of-the-art lacks research on whether the collaboration estimation models which are developed using multimodal data can generalize to different contexts (differs in learning activity, type of activity and schools). Third, the knowledge about what kind of features from different modalities enables the development of more generalizable collaboration estimation models is currently missing. Thus, the current paper tries to tackle those three research gaps.

STUDY SETUP

Context

The study was conducted in classrooms of two different types of schools in Estonia: vocational school and upper secondary school. In vocational school, the data were collected from collaborative learning activities in 6 classrooms with 4 different teachers. The subjects were mathematics, chemistry for woodwork (a chemistry course specifically designed for woodwork students integrating chemistry and woodwork), Estonian language and English language. Students used headphones primarily to use the microphones only as they were in F2F settings and did not have to listen to one another through headphones. In upper secondary school, there were two classroom sessions with 2 different teachers. The subjects were class-teacher lessons and biology. Students did not have headphones while the data was collected. Thus, this dataset had a lower quality of audio than the first dataset (i.e., from the vocational school). The classrooms, in general, had less than 30 students and the group sizes varied between two to four. The students were older than 18 years in both schools. Participants had prior experience working in collaborative groups. Additionally, an introductory session was provided to students before the activity, offering information on effective collaboration. The participants were of Estonian background and the languages for communication were Estonian and English (refer to [Table 1](#)). The group activity involved the use of a collaborative text editor. During the group activity, the teacher was present in the classroom and monitoring the groups. Additionally, the teachers were also closely monitoring specific groups in person and intervened if needed. The teachers involved did not have any formal training on implementing and supporting collaborative learning activities. The groups were formed by the teacher based on the number of available students. Refer to [Table A1](#) in the Appendix for details of learning activity tasks.

Data collection tool

For conducting the collaborative learning activity and data collection, we used a web-based application, CoTrack (Chejara, Kasepalu, et al., 2023). CoTrack uses an open-source collaborative text editor, Etherpad, to allow group members to draft the solution to a given task in collaboration. In addition, CoTrack also enables data recording as well as data pre-processing (e.g., computing speaking time, performing speech-to-text in real-time). We collected audio, video and log data from CoTrack. [Figure 1](#) shows one group of students working on the given task and students' collaborative space in CoTrack.

TABLE 1 Datasets.

School	Activity type	Dataset	Subject	Language	Groups	Students	Instances (30s windows)
School 1	Group discussion	Dataset-1	Maths	Estonian	4	13	429
		Dataset-2	Chemistry for woodwork	Estonian	5	15	344
		Dataset-3	Estonian language	Estonian	3	7	145
	Collaborative writing	Dataset-4	English language	English	2	6	210
		Dataset-5	English language	English	4	12	210
		Dataset-6	English language	English	2	4	209
School 2	Group discussion	Dataset-7	Class teacher lesson	Estonian	5	16	235
		Dataset-8	Biology	English	2	7	88

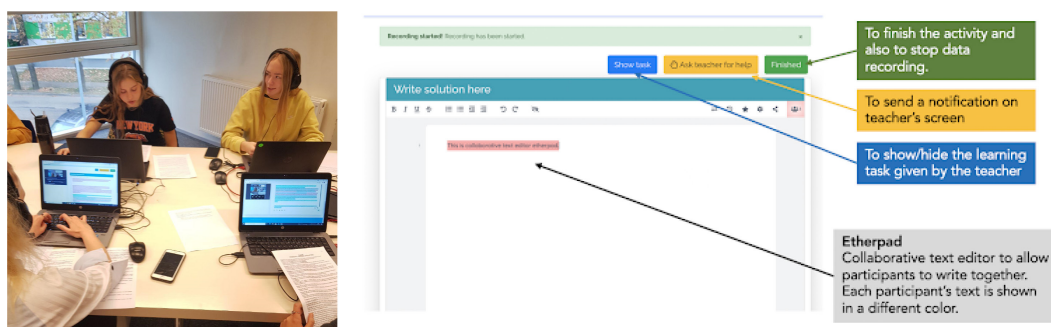


FIGURE 1 Study setup.

Procedure

A researcher (also co-author) co-designed learning activities with the concerned teachers before the study. The same researcher was also present in the classroom during data collection and provided a brief introduction to the study to the students. Following that consent was taken from the students. The teacher then grouped the students and the researcher provided instructions to start the group activity in CoTrack. The duration of the group activities varied from 20 to 60 minutes.

Data features

In CSQL scenarios, students share social–emotional cues while interacting with digital tools. Well-designed technology-mediated interactions facilitate the sharing and building of knowledge as well as the sharing of emotional states (Isohäätä et al., 2020). In these kinds of settings, the use of audio and video data has the potential to offer insights into collaborative learning processes. For example, the number of conversation turns, and speaking time have been found to be associated with collaboration quality in prior work (Chejara, Prieto, Rodriguez-Triana, Kasepalu, et al., 2023; Chejara et al., 2021; Martinez-Maldonado et al., 2013; Praharaj et al., 2021). Similarly, Storch, 2001 in their discourse analysis of collaborative pairs found differences in high/low collaborative groups in terms of their usage of personal pronouns. Based on these, we extracted speaking time, turn-taking and frequency of 'I', 'We' and Wh-words (e.g., what, why). The video data enables the extraction of facial action units which could potentially be used as a proxy for capturing students' emotional states (e.g., frustration) in educational settings (Craig et al., 2008). These states have a bi-directional relationship with learning, they can be a cause or a result of learning (Fiedler & Beier, 2014). Research has also identified facial expressions to be associated with collaboration dimensions (Hayashi et al., 2019). For example, in their study, Hayashi et al. (2019) found a strong association between anger and mutual understanding. Thus, we extracted facial action units from video data based on these aforementioned works. Additionally, we also extracted head orientation and mouth area region features. Head orientation provides an estimate of students' gaze which offers valuable insights into students' engagement and attentional processes (Thomas & Jayagopi, 2017). The mouth area region can also be used to detect student-generated speech (Siatras et al., 2009). We also computed features, i.e., characters written or deleted, from Etherpad logs, since they can offer insight into individual and group participation (Weinberger & Fischer, 2006). All the aforementioned features were extracted at the individual level and later their average and standard deviation were computed for group-level features. Refer to Table A2 in the Appendix for details on extracted features and their related studies in the field.

Collaboration quality ground truth

To annotate the quality of collaboration, we used the rating scheme from Rummel et al., 2011. This rating scheme specifies seven dimensions, namely, structuring problem-solving and time management, argumentation, cooperative orientation, knowledge exchange, collaboration flow, sustaining mutual understanding and individual task orientation. These dimensions were annotated for every 30-second time window on the 5-Likert scale from -2 to +2, following previous research works in MMLA (Martinez-Maldonado et al., 2013). These scores were added to get a final score for collaboration quality. Four MA students were trained to manually annotate video recordings with log data in two rounds. The interrater reliability score (Cohen's kappa = 0.61) for each dimension was above a substantial level.

METHODS

Correlation analysis

We performed a correlation analysis to identify relationships between multimodal data features and collaboration quality (and its dimensions). We used a non-parametric test (Spearman) because of normality assumption failure (Shapiro–Wilk test).

Model development

For model development, we decided to use a temporal window size of 60s for the segmentation of the dataset. This decision was based on our prior study which explored different window sizes (30s, 60s, 90s, 120s, 180s, 240s) and found that 60s window size enabled improvement in the model's performance across contexts (Chejara, Prieto, Rodriguez-Triana, Ruiz-Calleja & Khalil, 2023). The selected window size was used for the segmentation of the dataset. We then used the dataset from school-1 to configure different modelling pipelines (32 in total). These pipelines involved outlier detection, data scaling, hyper-parameter optimization, the use of contextual features (e.g., number of students, language of communication, type of learning activity, etc.) and threshold selection. We then identified high-performing modelling pipelines for our final model development (Chejara, Prieto, Rodriguez-Triana, Kasepalu, et al., 2023). We trained a Random Forest algorithm on the data. We used Random Forest because of its high performance for collaboration modelling tasks (Reilly & Schneider, 2019; Viswanathan & Vanlehn, 2018).

Model evaluation

We performed a model evaluation for 4 different levels of generalizability, namely, within context, across tasks, across task types and schools. These levels of generalizability are derived from an evaluation framework for assessing machine learning models in MMLA (EFAR-MMLA; Chejara et al., 2021). The framework specifies different levels of generalization and ways to assess them. For example, the first level is within context generalizability and assesses model performance on data instances coming from the same dataset within a particular context (but not seen during training). For this assessment, we used a 10-fold cross-validation (CV). Similarly, there are generalizability assessments performed across tasks where datasets differ on the given task or type of task. These two levels (difference in task and difference in type of task) were assessed using leave-one-dataset-out and leave-one-activity-out. For example, in

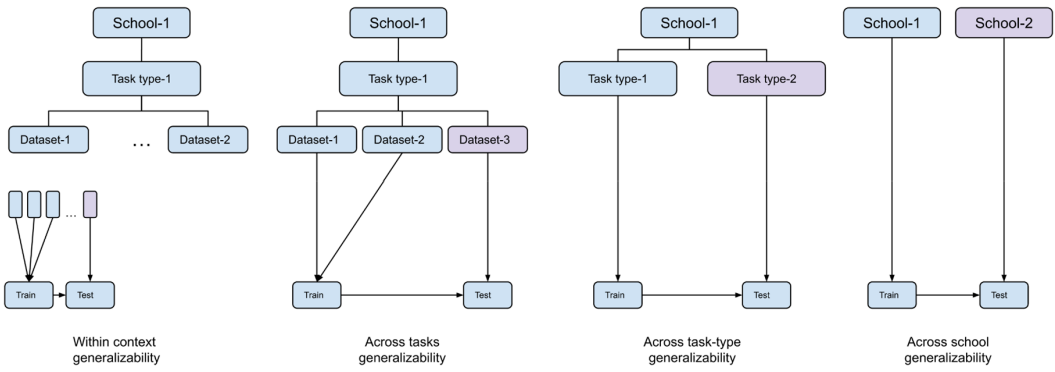


FIGURE 2 Generalizability evaluation at different levels.

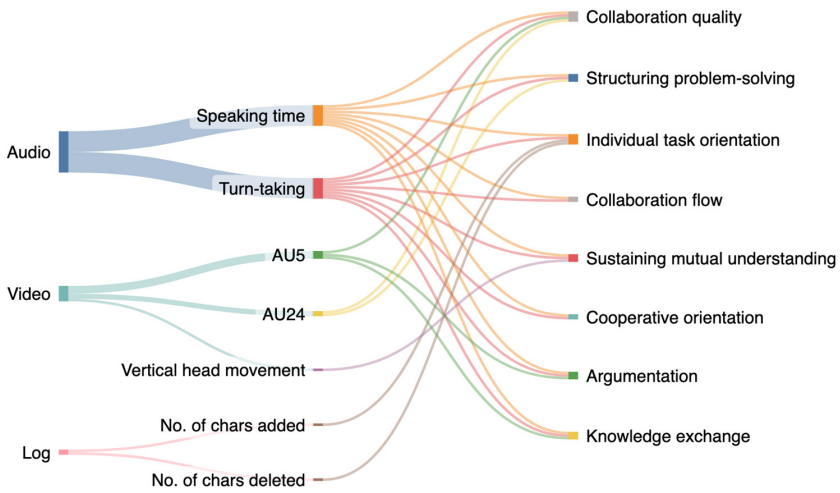


FIGURE 3 Relationship between multimodal data and collaboration quality (and its dimensions).

leave-one-activity-out (Figure 2), models were trained using datasets of collaborative-writing activity type and then assessed on datasets from the group-discussion type of activities and vice versa. Finally, we also assessed our model on a dataset from another school. Figure 2 shows generalizability levels and their assessment methods.

RESULTS

Relationship between multimodal data and collaboration quality in authentic classroom settings (RQ1)

We identified several multimodal data features that were found to be associated with collaboration quality and its underlying dimensions. Figure 3 provides a summary of the identified relationships:

Table 2 presents Spearman correlation measures between multimodal data features and collaboration quality and its dimensions. All the relationships found were positive, having weak to moderate correlations. Non-verbal features (speaking time, turn-taking)

TABLE 2 Spearman correlation measures between features from different modalities and collaboration quality dimensions.

Feature	Group level fusion	STR	ITO	CF	SMU	ARG	CO	KE	CQ (overall)
Frequency of AU05 (upper lid raiser)	Mean	–	–	–	–	0.27	–	0.28	0.25
	SD	–	–	–	–	0.25	–	0.26	–
Frequency of AU24 (lip presser)	Mean	0.26	–	–	–	–	–	–	0.25
Variation in head rotation (x-axis)	Mean	–	–	–	0.35	–	–	–	–
Average speaking time	Mean	0.32	0.30	0.32	0.35	0.28	0.33	0.36	0.36
	SD	–	–	–	–	–	–	0.25	–
Number of speaking turns	Mean	0.34	0.30	0.32	0.37	0.28	0.34	0.37	0.37
Total characters added	Mean	–	0.30	–	–	–	–	–	–
	SD	–	0.29	–	–	–	–	–	–
Total characters deleted	Mean	–	0.25	–	–	–	–	–	–
	SD	–	0.26	–	–	–	–	–	–

Note: Only the significant correlations are shown (p -value < 0.05) (STR: Structuring problem-solving and time management, ITO: Individual task orientation, CF: Collaboration flow, SMU: Sustaining mutual understanding, CO: Cooperative orientation, KE: Knowledge exchange, CQ: Collaboration quality).

were the only features that were found to be associated with collaboration quality and all seven dimensions. Most of those correlations were of moderate strength ($\rho > 0.30$) and the majority of the identified relationship for video features was found to be weak ($\rho \leq 0.30$). For example, we found facial action units AU05 (upper lid raiser) as positively correlated with argumentation ($\rho = 0.27$), knowledge exchange ($\rho = 0.28$) and overall collaboration quality ($\rho = 0.25$). Only head movement along the x-axis (moving head up and down) was positively correlated with sustaining mutual understanding dimension with a relatively stronger correlation ($\rho = 0.35$).

We also performed correlation analyses separately on datasets from school-1 and school-2. In both schools' data, audio features (speaking time, turn-taking) were positively correlated with collaboration quality and its dimensions for group-discussion-type activities. We also found the detected mouth region area—which was used as proxy for speaking activity (Siatras et al., 2009)—as positively correlated with collaboration quality and most of its dimensions in school-2's datasets. In both cases, the positive relationship between log features and individual task orientation remains the same. However, the relationship between AU24 and collaboration quality was inverted for school-2's datasets ($\rho = -0.31$) and shifted from weak to moderate strength.

Evaluation of collaboration quality estimation models across different levels of generalizability (RQ2)

Figure 4 shows the average balanced accuracy of the collaboration quality estimation models developed using audio, video and log data. The performance is reported on different levels of generalization, namely, within context, across collaborative-writing tasks (cr tasks), across group-discussion tasks (gd tasks), across different types of tasks

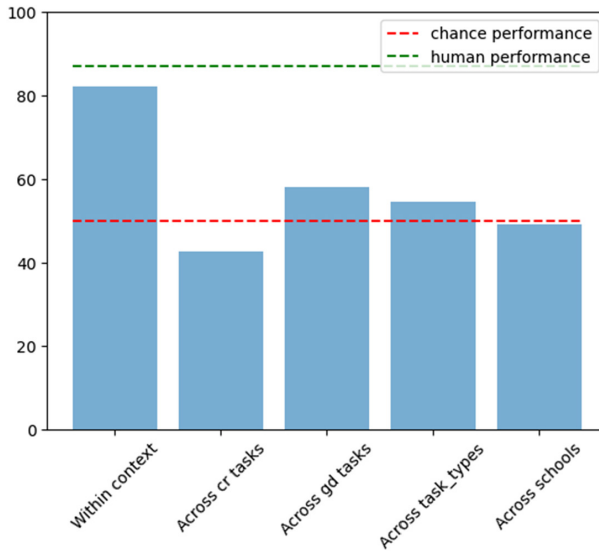


FIGURE 4 Balanced accuracy of random forest models for estimating collaboration quality using audio, video and log features at different levels of generalizability.

and across schools. The model achieved a performance of 82.0% (balanced accuracy), which was close to our human coders' performance (87.0%) at within context level. However, when the model was evaluated on different collaborative-writing tasks using leave-one-dataset-out, it performed worse than the chance model (42.5%). For group-discussion tasks, model performance dropped from 82.0% to 58.0%. The model then showed a further degradation of 3% at generalizability across different types of tasks. The performance was degraded to chance model's performance (~50%) for across-school generalization.

Comparative analysis of features from different modalities towards estimating collaboration quality in different contexts (RQ3)

Figure 5 presents the collaboration quality estimation models' performance at different levels of generalizability. The models in general performed well (above 70% balanced accuracy) when evaluating withincontext generalizability. The model with audio, video and log performed the best, achieving 82.0% balanced accuracy (Figure 5a).

When evaluating the generalizability across different collaborative-writing tasks (Figure 5b), the model did not achieve even chance performance and also had higher variation. The model with video data achieved performance closest to chance performance (49%). Regarding generalizability across group-discussion tasks (Figure 5c), models were able to better generalize. The audio-based model achieved the highest performance of 67.7%. In the case of generalizability across different types of tasks, audio-video-based models achieved the highest performance (59.0%, Figure 5d). For acrossschool generalizability (Figure 5e), audio modality emerged as the best-performing single modality (57.0%). Please refer to Tables A3–A7 in the Appendix for the model's performance in terms of other metrics (e.g., accuracy, precision, recall, kappa, f1-score).

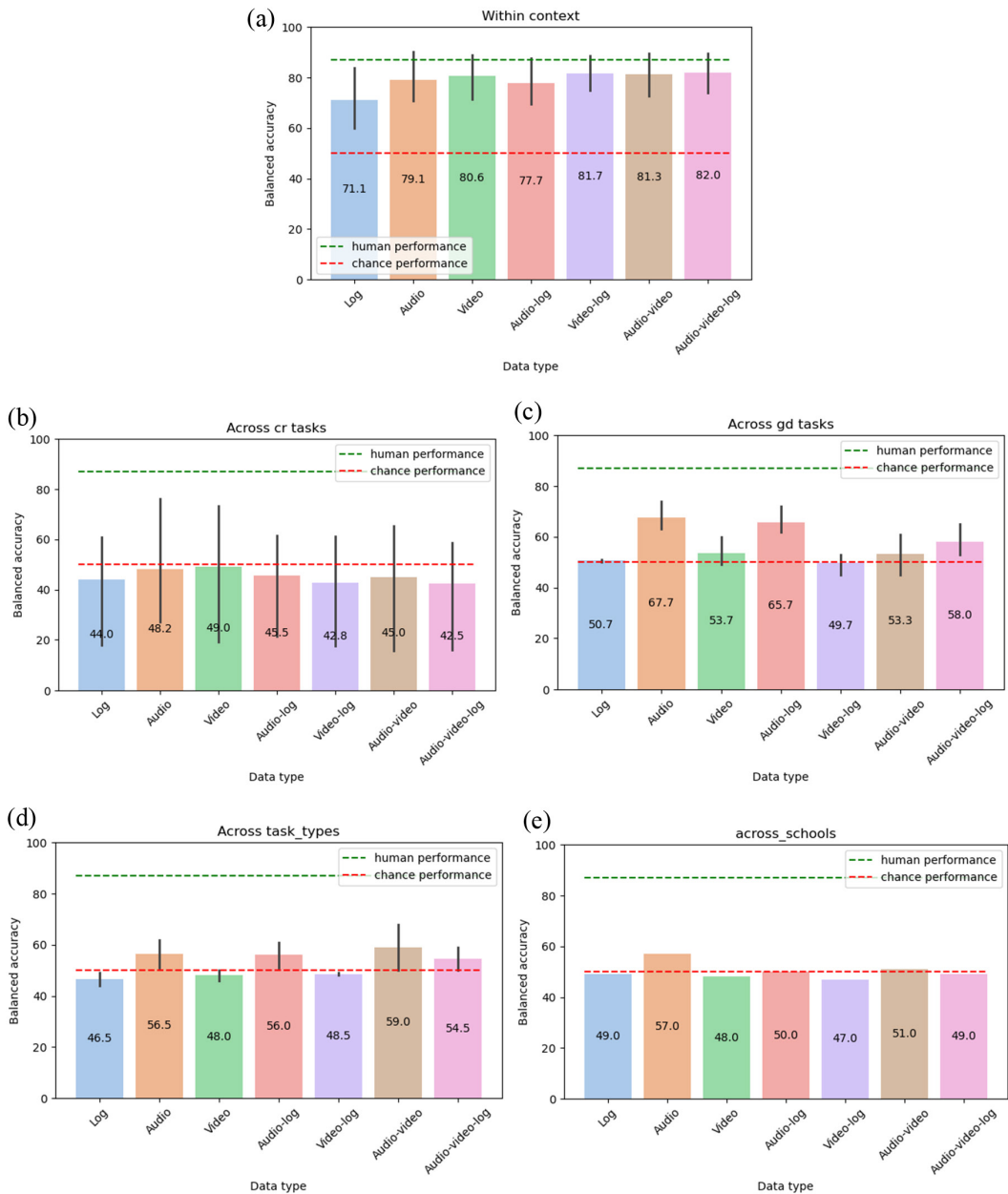


FIGURE 5 Balanced accuracy of random forest models for estimating collaboration quality using different modalities (with a 95% confidence interval).

DISCUSSION

RQ1: What is the relationship between multimodal data (audio, video and log) and collaboration quality (and its dimensions) in authentic classroom settings?

Finding 1: Speaking time and turn-taking are positively correlated with collaboration quality and its dimensions.

Speaking time and turn-taking features were found to have a positive correlation with collaboration quality and its dimensions. Even though the quality of audio was comparatively

lower in the school-2 dataset, our correlation analysis performed separately on datasets from both schools supported this finding regardless of the change in the data collection medium. The possible explanation is that the use of the laptop's inbuilt microphone in the school-2 dataset might have affected the computation of individual features, but it still provided some useful information at the group level (e.g., group's verbal participation).

This finding is aligned with prior work which found speaking time and turn-taking as predictive indicators of collaboration quality (Reilly & Schneider, 2019). For example, equal speaking time (i.e., each group's member participating equally, as measured by the Gini coefficient of group members' speaking time, with perfect equality being Gini = 0) was found to be associated with the overall quality of collaboration (Praharaj et al., 2021). This relationship could be explained by the nature of effective collaboration where communication plays a crucial role (Rummel & Spada, 2005). This includes students sharing their understanding, asking questions for clarification, offering feedback to others and supporting each other while working together (Webb, 2009). Thus, verbal participation plays an important role in effective collaboration. This behaviour can be partly captured through simple quantitative metrics of speaking time and turn-taking. This could explain the identified relationship between speaking time, turn-taking and collaboration quality.

Finding 2: Vertical head movement (nodding) correlates positively with sustaining mutual understanding.

Sustaining mutual understanding entails the group's members being on the same page in terms of problem understanding and solving it. Our finding on positive relationships between sustaining mutual understanding and vertical head movement (looking up and down) could be explained by head nods. Head nods are often used for acknowledging purposes in conversation to convey a signal of understanding (Duncan, 1972). This signal is likely to be used frequently in the group where group members support each other's ideas and maintain a common understanding of the problem and the worked-out solution. However, this finding needs further validation as cultural differences are likely to influence such social norms in group conversations.

Finding 3: AU05 (upper lid raiser) is positively correlated with argumentation, knowledge exchange and collaboration quality.

Facial action unit AU05 (upper lid raiser) is positively associated with collaboration quality and its dimensions of argumentation and knowledge exchange. The argumentation dimension entails students posing questions, discussing all available possible solutions and reaching a consensus. The positive relationship between AU05 and argumentation could be explained by previous research which investigated 16 personality traits using facial action units. Their study found AU05 associated with reasoning and warmth traits (Gavrilescu & Vizireanu, 2017). Other research studies have also found a positive association among AU05, attention and high arousal states (Frijda & Tcherkassof, 1997).

Finding 4: AU24 (lip presser) is positively correlated with structuring problem-solving and collaboration quality.

The positive relationship between lip presser (AU24) and structuring problem-solving can be explained by the emergence of negative emotions during collaborative learning activities (Cai et al., 2020). Structuring problem-solving involves students deciding strategies for group tasks which are likely to trigger negative emotions if all the students do not agree with the decision. Research studies have found a relationship between AU24 and anger (Sell et al., 2014). This suggests that a higher occurrence of lip pressers might indicate a tense situation during strategizing the group activity. The positive association between AU24 and collaboration quality may suggest that negative emotions (e.g., anger) can be associated with a high quality of collaboration. However, this relationship is in contrast with findings from our previous study where we analysed only a partial dataset using K-means clustering (Chejara, Prieto, Rodriguez-Triana, Ruiz-Calleja, Kasepalu, et al., 2023). In that study, we

identified the five most important features and one of those features was the lip presser action unit. This feature was found to have a negative relationship with collaboration quality, e.g., the cluster with higher values of AU24 was found to have lower collaboration quality scores. This may indicate that negative emotion might not be productive in every context.

Finding 5: Log features are positively correlated with individual task orientation.

The individual task orientation dimension looks at how motivated the student is towards solving the problem in the group. This could be also noticed in the participation level of the students in the group activities. Students with a high level of motivation tend to be more engaged with the task, thus, likely to have frequent interactions with the digital tool used in the learning activity. For example, previous research studies in learning analytics support this with their findings on the association between log features (e.g., reading time) and the motivation of students (Cocea & Weibelzahl, 2006). This could also explain the relationship between individual task orientation and the amount of students' writing captured in the form of a number of characters written or deleted.

RQ2: Whether and to what extent automated collaboration quality models generalize to different contexts varying on given tasks, the type of activities and school.

Finding 6: The models performed close to human performance when evaluated within the same contexts where data was gathered.

The findings suggest that models with audio, video and log features outperformed other modalities in terms of estimating collaboration quality within the context. This withincontext performance of a model in educational terms means that the model performs close to humans for data that come from the same context (same learning activity, the same groups of students, same teacher, classroom, etc). This finding is consistent with previous studies in MMLA which have reported high performance for collaboration quality classification tasks (Pugh et al., 2022; Viswanathan & VanLehn, 2018). However, we also like to mention the chances of overfitting the developed models during the withincontext evaluation due to a small dataset from each context.

Finding 7: The models did not generalize well for collaborative-writing type of activities.

For collaborative-writing type of activities, models performed worse than chance models. This poor performance can be explained by the nature of those activities (i.e., writing oriented) and the features used for modelling. This kind of task might not have generated much interaction among group members through face-to-face communication channels. This could have made it difficult for models to learn patterns from audio data. Plus, given our use of only simple log features, models might have been unable to learn collaboration patterns from those features. Thus, the use of more sophisticated log features could enable models to learn the differences between high- and low-quality collaboration, e.g., whether the group's members edit each other's writing or whether students are writing at the same time.

Finding 8: Automated model of collaboration quality generalizes to a different school with a 25% degradation in its performance.

The developed models showed a degradation of 25% in their performance (balanced accuracy from 82% to 57%) when used on datasets from another school. This finding has two implications: first, it offers mixed evidence on the generalizability of collaboration quality modelling using machine learning in MMLA; second, it suggests that the use of content-independent features may not provide a high-performing collaboration quality model for a higher level of generalizability. The often-reported high performance of collaboration quality models in MMLA so far has focused only on the lowest level of generalization (withincontext performance). For higher levels of generalization which is most likely to occur when the model is applied in authentic education settings, the models are likely to suffer from poor performance. However, we would like to emphasize that the goal of the developed models was not to offer an accurate estimation of collaboration quality, but rather to provide helpful cues to teachers regarding which group to visit and what to do there. Nevertheless, the use

of more sophisticated features (e.g., speech features using natural language processing) and information about pedagogical aspects (e.g., students' prior experience on group work) may help in improving the model's performance.

RQ3: What combination of multimodal data enables the development of a more generalizable collaboration estimation model?

Finding 9: Multimodality slightly improves performance when models are evaluated within context.

Multimodality is found to improve withincontext performance or in education terms, models estimated collaboration quality close to human performance in a particular context (with specific activity, specific students and specific subject). However, the multimodal nature of our data brought a slight improvement to the within context performance.

Finding 10: The model based on audio features achieved the highest performance for group-discussion-type activity.

Our machine learning model showed a higher performance for collaboration quality estimation tasks with audio data at across-task generalizability. In educational terms, this means that when models are used for estimation tasks on datasets coming from different contexts (different in terms of given task) then the use of audio data seems a better option than log or video data. This could be explained by tasks that were oriented more towards group-discussion than collaborative-writing. These tasks might have prompted students to interact more through face-to-face communication channels than with collaborative text editors. This explains why audio features enabled the development of high-performing collaboration quality models for group-discussion-type activities. This finding also provides evidence that audio data alone could help in building a more generalizable model for across-task contexts in *authentic* settings. This finding is consistent with prior research in MMLA (Pugh et al., 2022).

Finding 11: Audio–video modalities seem to be better for building models that are generalizable across different types of activities.

This finding suggests the use of audio–video features when building machine learning models for contexts where the type of collaborative learning activity is different. This could be explained by the differences in collaborative learning activities. For example, in our dataset, there were two different types of collaborative learning activities: group-discussion oriented and collaborative-writing oriented. The groups likely had more verbal interaction while working on group-discussion activities, versus the groups working on the collaborative-writing activities. Therefore, audio data alone might not be sufficient for models to learn generalizable patterns for both types of activities. The use of video data might have allowed models to learn non-verbal communication (e.g., head orientation).

Finding 12: Audio-based model performed better than the rest for across schools generalization.

Our results from across-schools generalizability evaluation showed that the audio-based model outperformed models using other modalities. This suggests that audio data alone can enable the development of collaboration estimation models that can be generalized across schools. This could be explained by the potential of audio data towards capturing social interaction data which provides important information for modelling collaboration. A prior study by Pugh et al., 2022 demonstrated it with their content-based model which was shown to generalize across different task domains (e.g., physical and math). The high performance of audio modality-based features also suggests that the use of additional modality does not necessarily bring improvement in model performance in every context.

Limitations

The presented research has six main limitations. First, the students were of Estonian background, therefore, the generalizability of findings needs further research to validate the

findings with students from different cultural backgrounds since this factor may condition the way students communicate. The second limitation is data noise caused by technical issues due to the authenticity of the setting. This led to the discarding of some groups' data from analysis which limited models to learn from a smaller set of groups rather than the entire classroom's data. The third limitation relates to our use of time-independent modelling methods only, while the collaboration process is likely to be time-dependent. The fourth limitation of the presented study is concerned with not using information about pedagogical aspects (e.g., students' understanding of effective collaboration). The fifth limitation is with our use of temporal windows. This can potentially support teachers in interventions with its estimation of collaboration quality. However, it is unlikely to support an assessment of overall collaboration quality for the entire duration. The final limitation is with the use of low-quality audio data from school-2. Though it allowed us to investigate the model's generalizability in authentic settings, it might also have affected the across-context generalizability evaluation when models were evaluated with school-2 dataset due to the lower quality of audio data.

CONCLUSION AND FUTURE WORK

This paper addresses a research gap in our understanding of the generalizability of automated models for collaboration quality estimation in authentic classroom settings. We developed machine learning models using audio, video and log data from a variety of collaborative tasks and subjects taking place at two different secondary/vocational schools in Estonia. Our results provide evidence of the use of speaking time and turn-taking features as potential collaboration indicators for *authentic* settings. Moreover, we also found vertical head movement as a potential indicator for sustaining mutual understanding across contexts. We also illustrated that models performed close to human performance (82% balanced accuracy) when evaluating within the same context of data gathering, but suffered a performance degradation of over 20% for across-school generalization. This could help the community to understand and assess the current state-of-the-art research on modelling collaboration using machine learning. In particular, the finding can help to comprehend the expected degradations in the performance of models which are developed in the field. This understanding could further help the community in identifying future research directions, for example, using context and content-related features (a cohesion matrix that provides information about the cohesiveness of spoken text) for collaboration modelling. In our future work, we plan to explore the use of those features for modelling collaboration quality using time-dependent analysis methods.

ACKNOWLEDGEMENTS

The presented work has been partially funded by the Estonian Research Council's Personal Research Grant (PRG) under grant number PRG1634. It also has been supported by grant RYC2021-032273-I, financed by MCIN/AEI/10.13039/501100011033 and the European Union's "NextGenerationEU/PRTR".

CONFLICT OF INTEREST STATEMENT

There was no conflict of interests in relation to the presented study.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ETHICS STATEMENT

The study presented in this paper has been approved by the Research Ethics Committee, Tallinn University, Estonia with application no. 6-5.1/24.

ORCID

Maria Jesús Rodríguez-Triana  <https://orcid.org/0000-0001-8639-1257>

REFERENCES

- Asterhan, C. S. C., Schwarz, B. B., & Gil, J. (2012). Small-group, computer-mediated argumentation in middle-school classrooms: The effects of gender and different types of online teacher guidance. *British Journal of Educational Psychology*, 82(3), 375–397. <https://doi.org/10.1111/j.2044-8279.2011.02030.x>
- Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences*, 12(3), 307–359.
- Cai, Y., Shimojo, S., & Hayashi, Y. (2020). Observing facial muscles to estimate the learning state during collaborative learning: A focus on the ICAP framework. In *ICCE 2020 - 28th International Conference on Computers in Education, Proceedings*, 1 (pp. 119–126). Asia-Pacific Society for Computers in Education.
- Chejara, P., Prieto, L. P., Ruiz-Calleja, A., Rodríguez-Triana, M. J., Shankar, S. K., & Kasepalu, R. (2021). EFAR-MMLA: An evaluation framework to assess and report generalizability of machine learning models in MMLA. *Sensors*, 21(8), 1–27. <https://doi.org/10.3390/s21082863>
- Chejara, P., Kasepalu, R., Prieto, L. P., Rodríguez-Triana, M. J., Ruiz-Calleja, A., & Shankar, S. K. (2023). Multimodal learning analytics research in the wild: Challenges and their potential solutions. In *Proceedings of the 6th Workshop on Leveraging Multimodal Data for Generating Meaningful Feedback (CROSSMMLA 2023) at the 13th International Learning Analytics & Knowledge* (pp. 36–42). <https://ceur-ws.org/Vol-3439/paper5.pdf>
- Chejara, P., Prieto, L. P., Rodríguez-Triana, M. J., Kasepalu, R., Ruiz-Calleja, A., & Shankar, S. K. (2023). How to build more generalizable models for collaboration quality? Lessons learned from exploring multi-context audio-log datasets using multimodal learning analytics. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (pp. 111–121). ACM. <https://doi.org/10.1145/3576050.3576144>
- Chejara, P., Prieto, L. P., Rodríguez-Triana, M. J., Ruiz-Calleja, A., Kasepalu, R., Chounta, I.-A., & Schneider, B. (2023). Exploring indicators for collaboration quality and its dimensions in classroom settings using multimodal learning analytics. In *Responsive and sustainable educational futures* (pp. 60–74). Springer, Cham. https://doi.org/10.1007/978-3-031-42682-7_5
- Chejara, P., Prieto, L. P., Rodríguez-Triana, M. J., Ruiz-Calleja, A., & Khalil, M. (2023). Impact of window size on the generalizability of collaboration quality estimation models developed using Multimodal Learning Analytics. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (pp. 559–565). ACM. <https://doi.org/10.1145/3576050.3576143>
- Chounta, I. A., & Avouris, N. (2012). Time series analysis of collaborative activities. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bio-informatics)* (Vol. 7493 LNCS(September), pp. 145–152). Springer. https://doi.org/10.1007/978-3-642-33284-5_13
- Chua, Y. H. V., Dauwels, J., & Tan, S. C. (2019). Technologies for automated analysis of co-located, real-life, physical learning spaces. In *LAK'19: Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, (pp. 11–20). Association for Computing Machinery. <https://doi.org/10.1145/3303772.3303811>
- Cocca, M., & Weibelzahl, S. (2006). Can log files analysis estimate learners' level of motivation? *Lernen, Wissensentdeckung Und Adaptivität, LWA*, 2006, 32–35.
- Craig, S. D., D'Mello, S., Witherspoon, A., & Graesser, A. (2008). Emote aloud during learning with AutoTutor: Applying the facial action coding system to cognitive—Affective states during learning. *Cognition and Emotion*, 22(5), 777–788. <https://doi.org/10.1080/02699930701516759>
- Di Mitri, D., Schneider, J., Specht, M., & Drachsler, H. (2018). From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning*, 34(4), 338–349. <https://doi.org/10.1111/jcal.12288>
- Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2), 283–292.
- Fiedler, K., & Beier, S. (2014). Affect and cognitive processes in educational contexts. In Pekrun, R., & Linnenbrink-Garcia, L. (Eds.), *International handbook of emotions in education* (pp. 36–55). Routledge.
- Frijda, N. H., & Tcherkassof, A. (1997). Facial expressions as modes of action readiness. In J. A. Russell & J. M. Fernandez-Dols (Eds.), *The psychology of facial expression* (pp. 103–132). Cambridge University Press.

- Gavrilescu, M., & Vizireanu, N. (2017). Predicting the sixteen personality factors (16PF) of an individual by analyzing facial features. *EURASIP Journal on Image and Video Processing*, 2017(1), 1–19. <https://doi.org/10.1186/s13640-017-0211-4>
- Gillies, R. M. (2019). Promoting academically productive student dialogue during collaborative learning. *International Journal of Educational Research*, 97(July 2017), 200–209. <https://doi.org/10.1016/j.ijer.2017.07.014>
- Goodman, B. A., Linton, F. N., Gaimari, R. D., Hitzeman, J. M., Ross, H. J., & Zarrella, G. (2005). Using dialogue features to predict trouble during collaborative learning. *User Modelling and User-Adapted Interaction*, 15(1), 85–134. <https://doi.org/10.1007/s11257-004-5269-x>
- Hadwin, A., & Oshige, M. (2011). Self-regulation, coregulation, and socially shared regulation: Exploring perspectives of social in self-regulated learning theory. *Teachers College Record*, 113(2), 240–264.
- Hayashi, Y. (2019). Detecting collaborative learning through emotions: An investigation using facial expression recognition. In *Intelligent Tutoring Systems: 15th International Conference, ITS 2019*, Kingston, Jamaica, June 3–7, 2019, Proceedings 15 (pp. 89–98). Springer.
- Hernández-García, Á., Acquila-Natale, E., Chaparro-Peláez, J., & Conde, M. (2018). Predicting teamwork group assessment using log data-based learning analytics. *Computers in Human Behavior*, 89(March), 373–384. <https://doi.org/10.1016/j.chb.2018.07.016>
- Huang, K., Bryant, T., & Schneider, B. (2019). Identifying collaborative learning states using unsupervised machine learning on eye-tracking, physiological and motion sensor data. In *EDM 2019 - Proceedings of the 12th International Conference on Educational Data Mining* (pp. 318–323). EDM.
- Huang, X., & Lajoie, S. P. (2023). Social emotional interaction in collaborative learning: Why it matters and how can we measure it? *Social Sciences and Humanities Open*, 7(1), 100447. <https://doi.org/10.1016/j.ssaho.2023.100447>
- Isohäätä, J., Näykki, P., & Järvelä, S. (2020). Cognitive and socio-emotional interaction in collaborative learning: Exploring fluctuations in Students' participation. *Scandinavian Journal of Educational Research*, 64(6), 831–851. <https://doi.org/10.1080/00313831.2019.1623310>
- Johnson, D. W., & Johnson, R. T. (1992). Key to effective cooperation. In R. Hertz-Lazarowitz & N. Miller (Eds.), *Interaction in cooperative groups. The theoretical anatomy of group learning* (pp. 174–199). Cambridge University Press.
- King, A. (2008). Structuring peer interaction to promote higher-order thinking and complex learning in cooperating groups. In *The teacher's role in implementing cooperative learning in the classroom* (pp. 73–91). Springer US.
- Liu, Y., Wang, T., Wang, K., & Zhang, Y. (2021). Collaborative learning quality classification through physiological synchrony recorded by wearable biosensors. *Frontiers in Psychology*, 12(674369). <https://doi.org/10.3389/fpsyg.2021.674369>
- Martinez, R., Kay, J., Wallace, J. R., & Yacef, K. (2011). Modelling symmetry of activity as an indicator of collocated group collaboration. In J. A. Konstan, R. Conejo, J. L. Marzo, & N. Oliver (Eds.), *User modeling, adaptation and personalization. Lecture Notes in Computer Science* (pp. 207–218). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-22362-4_18
- Martinez-Maldonado, R., Clayphan, A., Yacef, K., & Kay, J. (2015). MTFeedback: Providing notifications to enhance teacher awareness of small group work in the classroom. *IEEE Transactions on Learning Technologies*, 8(2), 187–200. <https://doi.org/10.1109/TLT.2014.2365027>
- Martinez-Maldonado, R., Dimitriadis, Y., Martinez-Monés, A., Kay, J., & Yacef, K. (2013). Capturing and analyzing verbal and physical collaborative learning interactions at an enriched interactive tabletop. *International Journal of Computer-Supported Collaborative Learning*, 8(4), 455–485. <https://doi.org/10.1007/s11412-013-9184-1>
- Martinez-Maldonado, R., Kay, J., Buckingham Shum, S., & Yacef, K. (2019). Collocated collaboration analytics: Principles and dilemmas for mining multimodal interaction data. *Human-Computer Interaction*, 34(1), 1–50. <https://doi.org/10.1080/07370024.2017.1338956>
- Ochoa, X. (2017). Multimodal learning analytics. In C. Lang, G. Siemens, A. F. Wise, & D. Gaevic (Eds.), *The handbook of learning analytics* (1st ed., pp. 129–141). Society for Learning Analytics Research (SoLAR).
- Olsen, J. K., Sharma, K., Rummel, N., & Alevin, V. (2020). Temporal analysis of multimodal data to predict collaborative learning outcomes. *British Journal of Educational Technology*, 51(5), 1527–1547. <https://doi.org/10.1111/bjet.12982>
- Ponce-Lopez, V., Escalera, S., & Baro, X. (2013). Multi-modal social signal analysis for predicting agreement in conversation settings. In *ICMI'13: Proceedings of the 2013 ACM International Conference on Multimodal Interaction* (pp. 495–501). Association for Computing Machinery. <https://doi.org/10.1145/2522848.2532594>
- Praharaj, S., Scheffel, M., Drachler, H., & Specht, M. (2021). Co-located collaboration modelling using multimodal learning analytics—Can we go the whole nine yards? *IEEE Transactions on Learning Technologies*, 14(3), 367–385. <https://doi.org/10.1109/TLT.2021.3097766>

- Pugh, S. L., Rao, A., Stewart, A. E. B., & D'Mello, S. K. (2022). Do speech-based collaboration analytics generalize across task contexts? In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 208–218). ACM.
- Reilly, J. M., & Schneider, B. (2019). Predicting the quality of collaborative problem solving through linguistic analysis of discourse. In *EDM'2019: Proceedings of the 12th International Conference on Educational Data Mining* (pp. 149–157). International Educational Data Mining Society (IEDMS).
- Rummel, N., Deiglmayr, A., Spada, H., Kahrimanis, G., & Avouris, N. (2011). Analyzing collaborative interactions across domains and settings: An adaptable rating scheme. In S. Puntambekar, G. Erkens, & C. Hmelo-Silver (Eds.), *Analyzing Interactions in CSCL*. Computer-Supported Collaborative Learning Series, 12. Springer. https://doi.org/10.1007/978-1-4419-7710-6_17
- Rummel, N., & Spada, H. (2005). Learning to collaborate: An instructional approach to promoting collaborative problem solving in computer-mediated settings. *Journal of the Learning Sciences*, 14(2), 201–241. https://doi.org/10.1207/s15327809jls1402_2
- Salomon, G., & Globerson, T. (1989). When teams do not function the way they ought to. *International Journal of Educational Research*, 13(1), 89–99.
- Schneider, B., Sung, G., Chng, E., & Yang, S. (2022). How can high-frequency sensors capture collaboration? A review of the empirical links between multimodal metrics and collaborative constructs. *Sensors*, 21(24), 8185.
- Sell, A., Cosmides, L., & Tooby, J. (2014). The human anger face evolved to enhance cues of strength. *Evolution and Human Behavior*, 35(5), 425–429. <https://doi.org/10.1016/j.evolhumbehav.2014.05.008>
- Sharma, K., Papavasopoulou, S., & Giannakos, M. (2019). Joint emotional state of children and perceived collaborative experience in coding activities. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children, IDC 2019* (pp. 133–145). Association for Computing Machinery. <https://doi.org/10.1145/3311927.3323145>
- Siatras, S., Nikolaidis, N., Krinidis, M., & Pitas, I. (2009). Visual lip activity detection and speaker detection using mouth region intensities. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(1), 133–137. <https://doi.org/10.1109/TCSVT.2008.2009262>
- Spikol, D., Ruffaldi, E., Dabisias, G., & Cukurova, M. (2018). Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *Journal of Computer Assisted Learning*, 34(4), 366–377. <https://doi.org/10.1111/jcal.12263>
- Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge*. MIT Press.
- Stiefelhagen, R., & Zhu, J. (2002). Head orientation and gaze direction in meetings. In *CHI 02 Extended Abstracts on Human Factors in Computer Systems CHI 02*, 1 (p. 858). Association for Computing Machinery. <https://doi.org/10.1145/506621.506634>
- Storch, N. (2001). How collaborative is pair work? ESL tertiary students composing in pairs. *Language Teaching Research*, 5(1), 29–53. <https://doi.org/10.1177/136216880100500103>
- Thomas, C., & Jayagopi, D. B. (2017, November). Predicting student engagement in classrooms using facial behavioral cues. In *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education* (pp. 33–40). Association for Computing Machinery.
- Viswanathan, S. A., & Vanlehn, K. (2018). Using the tablet gestures and speech of pairs of students to classify their collaboration. *IEEE Transactions on Learning Technologies*, 11(2), 230–242. <https://doi.org/10.1109/TLT.2017.2704099>
- Webb, N. M. (2008). Teacher practices and small-group dynamics in cooperative learning classrooms. In Gillies, R.M., Ashman, A.F., Terwel, J. (Eds.), *The teacher's role in implementing cooperative learning in the classroom* (pp. 201–221). Springer. https://doi.org/10.1007/978-0-387-70892-8_10
- Webb, N. M. (2009). The teacher's role in promoting collaborative dialogue in the classroom. *British Journal of Educational Psychology*, 79(1), 1–28. <https://doi.org/10.1348/000709908X380772>
- Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, 46(1), 71–95.
- Yoo, J., & Kim, J. (2014). Can online discussion participation predict group project performance? Investigating the roles of linguistic features and participation patterns. *International Journal of Artificial Intelligence in Education*, 24(1), 8–32. <https://doi.org/10.1007/s40593-013-0010-8>

How to cite this article: Chejara, P., Kasepalu, R., Prieto, L. P., Rodríguez-Triana, M. J., Ruiz Calleja, A., & Schneider, B. (2023). How well do collaboration quality estimation models generalize across authentic school contexts? *British Journal of Educational Technology*, 00, 1–23. <https://doi.org/10.1111/bjet.13402>

APPENDIX A

TABLE A1 Details of learning activities.

Dataset	Learning activity task	Subject	Duration
1	The collaborative activity involved solving a set of geometric problems. Each group was given a different set of problems. For example, one problem for group 3 was to calculate the perimeter and area of a rectangle with a diagonal of 8.5 dm forming an angle of 25 degrees with a larger side	Mathematics	60 minutes
2	The task involved a hypothetical situation of a person, Steve, who needed to renovate a particular portion of his house (exterior facade, bathroom and room). The groups were given a map of the house with measurements of each wall as well as the floor. The groups were asked to first prepare a list of tools and materials needed to complete the renovation. The groups were also asked to discuss the estimated cost of labour and materials, and prepare the final document with all details for Steve	Chemistry with woodwork	60 minutes
3	The task involved preparing a presentation in the group on one of the epic ^a topics (eg. Gilgamesh, Song of my Cid). The groups were given instructions on the content to put in the presentation, eg, describe the main characters and summarize the central story of the epic. At the end of the session, the groups were asked to present in front of the class	Estonian language	45 minutes
4	The task was to complete the given sentences on past and present tenses. The activity also asked groups to discuss and write collaboratively a paragraph on what they would do if they were given a particular sum of money (10,000 euros)	English language	50 minutes
5	The activity involved dividing student groups into two categories: Employee and Employer. For each category, students were given a set of questions/tasks to discuss and write down in the text editor. For example: one of the tasks for the Employer group was "You are the owners of a construction company, please think about which personal traits are important for a construction worker. Put down the traits below and also the reason why they are important"	English language	30 minutes
6	The task was to write an essay collaboratively on given topics (eg, the generation today is less healthy than our parents). The groups were also asked to assess their essay against a set of checklists focusing on content, communication, organization and language use	English language	60 minutes
7	The groups were given worksheets that had questions on DNA sequencing and mutation (eg, on the effect of GTA → GTT mutation) to be answered in the collaborative text editor	Biology	30 minutes
8	The students were given the task of planning a class trip involving collaboratively selecting a destination, allocating a budget (travel, meals and accommodation) and creating a schedule for the entire trip	Class teacher lesson	20 minutes

^aOxford definition: a long poem about the actions of great men and women or a nation's history.

TABLE A2 Extracted data features and their related studies.

Data type	Feature name	Description	Related studies
Audio	speaking_time	Speaking time in seconds	MMLA research studies have found speaking time and turn taking as indicators of collaboration (Martínez-Maldonado et al., 2013; Ponce-Lopez et al., 2013; Praharaj et al., 2021)
	turn_taking	Number of speaking turns taken by the participant	
	freq_l, freq_you & freq_we	Frequency of 'I', 'You' and 'We'	Storch (2001) found differences between high- and low- collaborating groups in terms of their use of personal pronouns
	freq_wh	Frequency of wh-words (i.e., what, why, who)	Question-asking behaviour indicates the argumentative nature of collaboration as per the rating scheme of Rummel et al. (2011)
Video	AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU11, AU12, AU14, AU15, AU17, AU18, AU20, AU23, AU24, AU25, AU26, AU28, AU43	Facial action units	Previous research found a relationship between facial action units and collaboration behaviour (Cai et al., 2020; Sharma et al., 2019)
	head_rotation_x head_rotation_y head_rotation_z	Head pose (head rotation along x-, y-, z-axis) as proxy for eye-gaze	Head pose contributes 68% to eye-gaze (Stiefelhagen & Zhu, 2002)
	mouth_area	Mouth region area as proxy for speaking activity	(Siatras et al., 2009)
	chars_add chars_del	Number of characters added Number of characters deleted	Indicators of individual participation have been found as one of the key quantitative metrics for collaborative learning (Weinberger & Fischer, 2006)

TABLE A3 Within context performance.

Modality	Accuracy	Balanced accuracy	Precision	Recall	Kappa	F1-score
Log	78 (17)	71 (18)	59 (34)	59 (38)	0.30 (0.25)	58 (36)
Audio	84 (10)	79 (13)	68 (32)	68 (34)	0.50 (0.21)	67 (33)
Video	88 (7)	80 (12)	73 (32)	75 (35)	0.51 (0.18)	73 (33)
Audio–log	84 (11)	77 (13)	68 (32)	68 (34)	0.44 (0.18)	67 (33)
Video–log	89 (6)	81 (10)	73 (33)	76 (35)	0.54 (0.16)	74 (33)
Audio–video	88 (6)	81 (12)	74 (33)	75 (35)	0.53 (0.18)	74 (33)
Audio–video–log	88 (7)	82 (12)	73 (33)	75 (34)	0.55 (0.19)	73 (33)

TABLE A4 Across task contexts (collaborative writing activities).

Modality	Accuracy	Balanced accuracy	Precision	Recall	Kappa	F1-score
Log	63 (40)	44 (26)	65 (44)	69 (46)	0.14 (0.15)	67 (45)
Audio	61 (36)	48 (29)	65 (44)	64 (45)	0.14 (0.24)	64 (44)
Video	56 (33)	49 (30)	66 (45)	59 (39)	0.10 (0.08)	62 (41)
Audio–log	64 (37)	45 (24)	65 (44)	68 (46)	0.14 (0.16)	67 (45)
Video–log	49 (39)	42 (25)	64 (43)	50 (47)	0.06 (0.10)	52 (45)
Audio–video	52 (36)	45 (29)	66 (44)	54 (38)	0.09 (0.09)	59 (41)
Audio–video–log	48 (34)	42 (26)	65 (44)	49 (37)	0.04 (0.07)	55 (39)

TABLE A5 Across task contexts (group-discussion activities).

Modality	Accuracy	Balanced accuracy	Precision	Recall	Kappa	F1-score
Log	55 (1)	50 (0)	38 (1)	31 (4)	0.02 (0.01)	34 (2)
Audio	71 (5)	67 (5)	67 (11)	52 (14)	0.37 (0.10)	57 (10)
Video	60 (0)	53 (5)	41 (9)	26 (28)	0.06 (0.09)	26 (21)
Audio–log	69 (5)	65 (5)	62 (11)	51 (13)	0.33 (0.09)	55 (9)
Video–log	56 (5)	49 (4)	34 (9)	20 (20)	-0.01 (0.07)	23 (16)
Audio–video	60 (7)	53 (8)	45 (16)	25 (23)	0.07 (0.13)	28 (18)
Audio–video–log	64 (2)	58 (6)	56 (4)	31 (26)	0.17 (0.09)	36 (18)

TABLE A6 Across task types.

Modality	Accuracy	Balanced accuracy	Precision	Recall	Kappa	F1-score
Log	47 (14)	46 (3)	40 (10)	12 (2)	-0.06 (0.03)	18 (4)
Audio	64 (4)	56 (7)	57 (19)	45 (48)	0.15 (0.12)	46 (40)
Video	49 (18)	48 (2)	21 (29)	3 (4)	-0.03 (0.03)	5 (7)
Audio–log	63 (2)	56 (7)	57 (18)	42 (48)	0.12 (0.10)	43 (42)
Video–log	50 (15)	48 (0)	34 (32)	6 (9)	-0.01 (0.01)	11 (14)
Audio–video	64 (3)	59 (12)	56 (32)	33 (45)	0.17 (0.17)	36 (49)
Audio–video–log	60 (2)	54 (6)	35 (49)	29 (41)	0.09 (0.09)	32 (45)

TABLE A7 Across schools performance.

Modality	Accuracy	Balanced accuracy	Precision	Recall	Kappa	F1-score
Log	51	49	42	24	-0.02	28
Audio	53	57	72	32	0.13	44
Video	57	48	26	22	-0.05	24
Audio-log	43	50	66	26	0.01	37
Video-log	56	47	26	22	-0.06	24
Audio-video	58	51	36	27	0.02	31
Audio-video-log	57	49	33	25	-0.02	28