Harbingers of Collaboration? The Role of Early-class Behaviors in Predicting Collaborative Problem Solving

Paul Hur
University of Illinois at Urbana-Champaign
khur4@illinois.edu

Nigel Bosch
University of Illinois at Urbana-Champaign
pnb@illinois.edu

Luc Paquette
University of Illinois at Urbana-Champaign
lpaq@illinois.edu

Emma Mercier
University of Illinois at Urbana-Champaign
mercier@illinois.edu

ABSTRACT

Collaborative problem solving behaviors are difficult to identify and foster due to their amorphous and dynamic nature. In this paper, we investigate the value of considering early class period behaviors, based on small group development theory, for building predictive machine learning models of collaborative behaviors during problem solving. Over 12 weeks, 20 small groups of undergraduate students solved problems facilitated by a digital joint problem space tool on tablet computers, in the 50-minute discussion component of an engineering course. We annotated 16,270 video clips of groups for collaborative behaviors including task relatedness, talk content, peer interaction, teaching assistant interaction, and tablet usage. We engineered two subsets of features from tablet log file data: onset features (early collaborative problem solving behavior characteristics calculated from the first ten minutes of the class) and concurrent features (more general collaborative behaviors from the whole class period). We compared accuracy between the onset, concurrent, and onset + concurrent features in machine learning models. Results exhibited a U-shaped pattern of accuracy over class time, and showed that onset features alone could not be used to effectively model groups' collaborative behaviors over the entire class time. Furthermore, analysis did not show support for significant gain in accuracy when onset features were combined with concurrent features. Finally, we discuss implications for studying collaborative learning and development of software to facilitate collaboration.

Keywords

Collaborative Problem Solving, Computer-Supported Collaborative Learning, Predicting Collaboration, Small Group Development

1. INTRODUCTION

Collaborative problem solving consists of the communication and coordination of shared effort between team members toward a common desired goal [19, 23, 26]. Though it has been identified as a critical skill for students in the classroom [11, 34, 25], it is difficult to identify effective behaviors and nurture them, since the nature of collaboration and teams can be amorphous [48] and dynamic [43]. Education and learning sciences researchers have advocated for qualitative coding of video data as a means to understand the complexities of learning behaviors [24], and have applied these methods to study collaborative behaviors and the development of collaborative practices in courses [35]. Computers can further support collaborative learning research through collaborative learning software, collaborative games, and digital joint problem spaces—"a socially-negotiated set of knowledge elements, such as goals, problem state descriptions and problem solving actions" [44]—the resulting log data of which have been widely used with machine learning and data mining approaches to uncover hidden patterns of collaborative behaviors [31, 1, 37, 12]. Recent technological advances have also given way to multimodal approaches, using eye-gaze tracking, bodily motion, and physiological data to identify collaborative states [28, 40].

Despite such diverse approaches to detect and identify collaborative behaviors in learning contexts, the evolution of collaborative practices in student groups has not been closely investigated. Understanding the evolution of collaboration and its impact on methods for measuring collaboration is crucial, however. What constitutes collaborative behaviors may change throughout a learning session [10], and thus measurement may need to be adapted as well. In this paper we focus on the relationship between measurement and behavioral changes over time within classroom sessions. In particular, we leverage organizational theory about the sequential nature of small group development to inform research on how to measure and predict collaboration via machine learning in the presence of inevitable shifts in behaviors throughout collaboration stages.

The rest of the paper is organized as follows: we first discuss the small group development theories on structured, sequential group development which motivated our work, then relate

them to collaborative problem solving in the classroom to define our research questions and respective hypotheses. We then introduce the context of our study, including the collaboration tool, behavior coding, data processing, and model building. Next, we present our findings and close with an interpretation of our results and note limitations and future work.

1.1 Small Group Development

Research in organization and management fields on understanding how collaborative behaviors contribute to small group dynamics and development goes back several decades. Perhaps most notably, Tuckman's 1965 meta-analysis of therapy and human relations training groups presented the *forming-storming-norming-performing* (and later a fifth stage, *adjourning* [50]) model [49]. The model outlined the existence of a sequential, stage-based trajectory of small group collaboration, in which a group must fulfill one stage before advancing to the next. Tuckman's five model stages were described as (1) orientation to task (forming), (2) emotional response to task demands (storming), (3) open exchange of relevant interpretations (norming), (4) emergence of solutions (performing), and (5) separation (adjourning). This has led to decades of efforts to better understand the stages in various settings, including management [36], education [51], and medical training [47].

Tuckman's 5-stage structure of group development was further supported by Cassidy's 36-book meta-framework study, which aimed to clarify group development for practical use by examining group development in therapy, education, and management settings [17]. Though some scholars have presented theoretical models with more or fewer stages to group development [46, 21, 54], others have supported the five-stage model with differently termed, but analogous stages to Tuckman's model [16, 22, 8].

In nearly all proposed theoretical models of small group development, the first stage is defined as the task orientation stage [49, 16, 22]. During this stage, group members contextualize the task within the given parameters and communicate regarding the manner in which it will be accomplished [49]. While "ground rules" are set during this stage, communication about task orientation continues on some level throughout the collaboration process. Moreover, in problem solving, communication with references to others' ideas rather than independent solution paths has been identified as an important marker of shared task alignment, or "establishment of a collaborative orientation toward problem solving" [4]. In this study, we briefly analyze transitions across the stages of small group development during problem solving in classrooms. However, we focus much more closely on the first stage, orientation to task, since it has been shown to have a significant positive effect on achievement [45]. The first stage characterizes cooperative orientation and the motivation to collaborate, which has a strong relation to the quality of collaboration [13].

1.2 Contributions and Novelty

This paper considers the role of early group behaviors in collaborative problem solving. We investigated whether explicitly incorporating early group behaviors as features improves machine learning predictions of collaboration and analyze how model accuracy evolves across time and stages of collaboration.

We used qualitative coding of collaborative behaviors on video data to measure collaboration. We then predicted those behaviors from features extracted from the action log files of a digital collaboration tool (run on tablet computers) used by undergraduate students in an introductory mechanical engineering course at a large Midwestern U.S. research university. We created various feature subsets and built corresponding machine learning models to evaluate the predictive accuracy of early group behaviors versus behaviors from later on in class periods. Assuming the presence of sequential, evolving collaborative behaviors in small groups, and the importance of early collaborative behaviors, machine learning models created from considering class behaviors as a whole may potentially be improved by accounting for early behaviors. For example, a group of students who fail to form a successful collaborative dynamic early on may struggle throughout class, whereas a group of students who exhibit high collaboration early on may be more effective in later stages. Consequently, we analyze whether a model built on features from class behaviors as a whole would have variable performance for collaborative behaviors predictions over the different segments of the class period, which align with the different stages of group development.

We aim to understand how effective collaborative behaviors, relating to orientation to task, during earlier stages may influence a group's collaborative behaviors in the future. As such, we also compare the performance a model solely built from such earlier features with one built from features of behaviors from all current and past in-class behaviors, not just early-stage behaviors.

We approach the aim of this paper by formulating and addressing several research questions:

RQ1 How does the predictive accuracy of collaborative behaviors vary across different periods of a 50-minute class?

Hypothesis: We expect stages of collaboration that are dominated by tablet computer interaction behaviors (e.g., reading, drawing) will be more successfully predicted than those dominated by discussion, and that the changing base rates of collaborative behaviors over time will influence classification accuracy [29].

RQ2 Can early class collaborative behaviors alone be used to effectively model and predict collaborative behaviors of the entire class period?

Hypothesis: We expect early class behaviors to predict the quality of collaboration later in class if and only if groups' collaboration quality remains static or consistently mirrors early collaboration.

RQ3 Are collaborative behavior prediction models improved through emphasizing early class collaborative behavior features?

Hypothesis: We expect prediction models will be more accurate later in class periods if early class behaviors capture

groups that are consistently collaborative or consistently not collaborative.

2. RELATED WORK

In this study, we utilized video coding methods along with machine learning approaches for analyzing action log data to study temporality. Work in computer-supported collaborative learning (CSCL) has highlighted the importance of considering temporality in collaboration, and Reimann has argued that "the main object of analysis in CSCL is a process—something that unfolds over time" [42]. For example, Mercier et al. examined through video coding and counting how the development of collaborative practices in engineering courses evolve over four weeks, and saw that patterns of interactions, such as conversation and workflow, change over time [35]. Others highlighted the value of utilizing more complex quantitative methods over video coding methods to consider temporality in analyzing problem-solving processes in computer-supported collaboration settings, as it can reveal aspects of group interactions that coding methods cannot reveal [32].

Collaborative learning may also be effectively analyzed via the action logs, discourse data, and gameplay data of digital tools and serious games, which are able to provide fine-grain recollections of the learner's interactions with the respective software. Educational data mining researchers have applied supervised [41] and unsupervised [14, 31] machine learning techniques to better understand collaboration and to inform the design of interventions to support collaborative learning through such means as software prompts [31] and content creation suggestions [52]. Additionally, Paquette et al. have highlighted the need to support students during collaborative learning by considering the role of the instructor in facilitating student collaboration [38]. As such, instructor dashboards have been explored as ways for instructors to more easily gauge and analyze student collaboration across multiple groups [3, 33]. A central aim of our study has been to inform better instructor interventions for facilitating collaboration through insights gained from analysis of action log data.

3. METHODS

This study utilizes data collected from a design-based implementation research project which aims to better facilitate collaboration in engineering problem solving through the analysis of video and interactions from engineering classes. The project team has developed a student-facing tool that facilitates student group collaboration through a synchronized-per-group shared digital environment (Figure 1) on tablet computers, which group members can use to create and display their work. During use, we collected two types of data: student interactions on the tool stored in log files—detailing actions taken by individual students such as writing, drawing, or editing—and video data from cameras set up around the classroom. One of the key goals of the tool is to scale to large classrooms where cameras are unlikely to be consistently available; thus, we utilize video data to collect ground truth labels, but rely only on logged tablet actions for collaboration prediction.

Data in this study came from the use of the tool in Fall 2017 during the discussion component of an undergraduate introductory mechanical engineering course at a large Midwestern U.S. research university. The research team worked closely with faculty and teaching assistants (TAs) to design tasks suitable for collaboration and in line with the intended learning outcomes from the class. The tasks were independent from week-to-week and did not build on one another, and the students were not graded on completion by the end of each class period. The tasks were represented in the tablet tool as worksheets with variable number of pages, which included problem descriptions and space to work out solutions. Data were collected across 12 weeks of class from 20 groups of approximately 4 students (group sizes varied from week to week based on attendance).

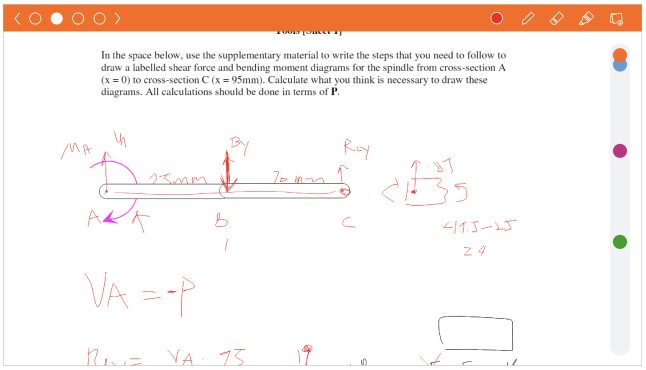

Figure 1: One example of the result of collaborative problem solving through the tool's shared digital environment. The interface allows students choices of different colors and tools to write, draw, and create figures.

While students interacted on tablets using the interface shown in Figure 1, TAs present in the classroom viewed student progress on their own tablets (Figure 2). The TA tablets showed students' editing positions in the worksheets, and allowed TAs to join any group as a non-interactive participant to see students' work in detail. Our current work seeks to augment the TA-facing tool via predictions of various markers of collaboration quality made by machine learning models. This feature enables TAs, who may lack extensive training in assessing and promoting collaboration, to identify groups that are not collaborating well and intervene to encourage collaboration.

3.1 Behavior Coding Process

Videos of each group's interactions (Figure 3) were captured by high-angled cameras and synchronized with audio data captured by microphones positioned near each group; additionally, an overhead fisheye lens camera captured the entire class, including events such as the TAs' interactions with groups. The collected video data were annotated (coded) at the group level by two trained annotators with an annotation scheme adapted from previous work on collaborative behavior annotation [38] to define group activity in terms of task relatedness, peer verbal interaction, TA interaction, talk content, and tablet usage.

Previous research on predicting collaboration from interactions with software has involved annotating similar content

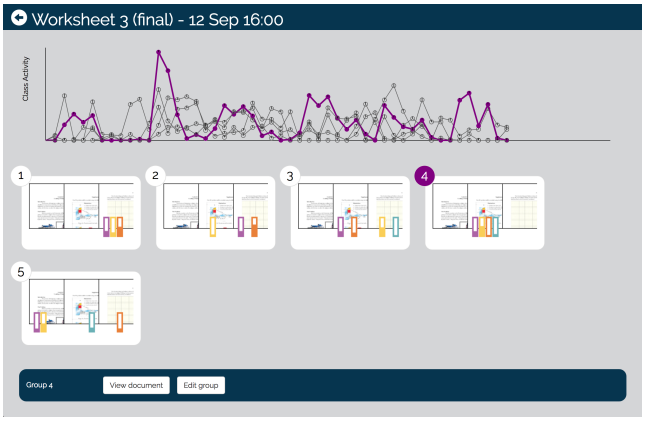

Figure 2: Example screenshot of a teaching assistant's view of a classroom with five groups of students. The top graph indicates activity over time for each group, with the selected group (#4) highlighted in purple. Bars on each worksheet thumbnail show the page each student is viewing and their individual levels of activity.

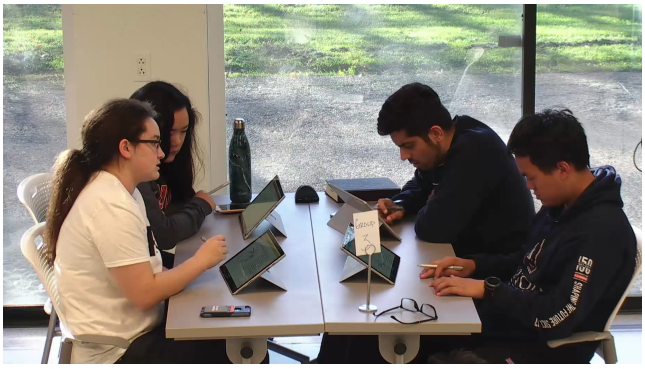

Figure 3: An example of a group working collaboratively on a problem through the tool on tablet computers. Videos of such groups were recorded and qualitatively coded through a coding scheme adapted from Paquette et al.'s work [38].

in video clips at 60-second intervals [38]. In this study, the presence of collaborative behaviors (expanded below) were annotated at 20-second video clips, after trials of the annotation process at different clip duration of 10, 20, 30, 40, and 60 seconds. Annotators determined 20 seconds to be a reasonable balance—10 seconds was too brief to confidently observe the presence of collaborative behaviors, while 30 seconds was too long and often led to the observance of multiple collaborative behaviors within the same video clip. Furthermore, through our trials at varied clip lengths, additional identifiable behaviors emerged that were better identified at the current 20-second coding clip length rather than the longer 60-second clips annotated in previous work. A total of 16,270 clips were annotated for the presence (annotated as 1), or absence (0) of the following set of collaborative behaviors:

- *Task relatedness*: At least one of the group members appears to be on task (e.g. two students solving problems on the tablet).
- *Peer verbal interaction*: Verbal interaction is present between group members.
- *TA class interaction*: TA is talking to the whole class (e.g., class-related announcement, addressing a frequently asked question).
- *TA group interaction*: TA is verbally interacting with at least one of the group members.
- *Task talk*: Audible talk content in the group is related to solving the task.
- *Other talk*: Audible talk content in the group is not related to solving the task.
- *Tablet movement*: At least one of the group members is moving the tablet to initiate (and to end) sharing of the screen content with others.

We measured inter-rater reliability via Cohen's kappa [18] and percent agreement on a subset of 2,125 video clips. Table 1 shows these reliabilities. All labels except *Other talk* (kappa = .651) achieved kappa = .8 or higher, indicating substantial agreement [18]. Given this agreement, the two annotators divided the remaining 14,145 clips and annotated them individually.

Table 1: Inter-rater reliability for a sample of 2,125 video clips in this study.

Behavior	Base rate	Agreement	Kappa
Task relatedness	.954	98.6%	.840
Peer verbal interaction	.501	91.7%	.833
TA class interaction	.024	99.5%	.898
TA group interaction	.150	98.3%	.932
Task talk	.608	91.2%	.816
Other talk	.072	95.3%	.651
Tablet movement	.019	99.2%	.801

Of the qualitatively coded behaviors, we considered six specific behaviors for this study: ON-TASK (derived directly from *Task relatedness*), ON-TASK-NO-INTERACTION (from a combination of *Task relatedness* and *Peer verbal interaction*), PEER-INTERACTION (from *Peer verbal interaction*), SILENT (from *Task talk* and *Other talk*), TASK-TALK (from *Task talk*), and TA-CLASS (from *TA class interaction*). These six behaviors were those which we believed would be best suited for investigating the evolution of collaboration with consideration of the actions of both the TA and students during a typical class period for the course. We did not include tablet movement due to the low base rate during annotation and questionable value for characterizing collaboration. We deemed ON-TASK-NO-INTERACTION important for distinguishing collaboration from individual work, and while it was not explicitly annotated, it was calculated from a combination of two different behavior labels (*Task relatedness* and *Peer verbal interaction*).

3.2 Data Processing

The tablet tool collected student action log files, one per group, during each class session. Relevant behavior data were cleaned and stored based on expected suitability for predicting collaborative behaviors on the tool. These types of data included event types, such as scrolling, drawing, object creation (inserting one of a few built-in graphics), modifying drawings or objects (removing or undoing), as well as the size and position of edits made, object geometry changes (e.g., moving, resizing), page number, scroll bar position, and changes to drawing color.

3.3 Machine Learning Models from Feature Subsets

We aligned annotated behaviors with the student action log-files to allow synchronized analysis between the two data sources. We created three features sets: (1) *onset* features, which characterized collaborative behaviors found in and calculated within the first ten minutes, (2) *concurrent* features, which captured collaborative behaviors based on the most recent 60 seconds as well as all cumulative data, and (3) *combined* features, which combines both subsets. Student behaviors were recorded individually within each group's log file. However, we primarily extracted features intended to characterize whole-group behaviors, in line with the group-level video annotation scheme and the overall project goal of improving collaboration rather than individual learning behaviors.

3.3.1 Feature Engineering

Designing features to extract took place over the course of several sessions involving the video annotators and researchers, who discussed behaviors observed in the classroom and how they might be reflected in tablet-based behaviors.

We extracted 89 features from the action logs using the full 50-minutes of the class duration, which we refer to as concurrent features. For these features, we used a combination of the behaviors that annotators had observed to be related to collaboration, as well as those characteristics we hypothesized to be more broadly associated with effective collaboration. For example, we created features such as: the mean distance between consecutive edits of the same students (since it may distinguish working in one area vs. jumping around rapidly), total number of unique document pages viewed (a higher number may symbolize more exploration of the task), and maximum distance between concurrent edits of the same page but made by different students (may symbolize task division).

Similarly, we extracted 21 features from the action logs calculated from the first ten minutes of class, which we refer to as onset features. Assuming the five stages of small group development apply in this context, we approximately split each 50-minute class period into stages by dividing into fifths. We expected each class period to somewhat reset the collaboration process, since there was a new task each week—meaning a new corresponding task identification stage (storming), as well as some variation in the group, in number and person, due to fluctuating attendance. We specifically kept in mind the characteristics of the task identification stage, such as verbal and written communication for contextualizing the problem and setting "ground rules", as well as behaviors such as reading or using visual figures to understand (but not necessarily solve) the exercises. To that end, we created features such as: the proportion of the first ten tool objects created by the group being the pre-made available diagrams (a higher proportion may mean more complete solutions early in class), the longest time between object additions and edits (longer pauses between actions may characterize more verbal communication), and the cumulative number of page switches (switching back and forth between pages may signal wanting to fully understand the task at hand by referencing material on other pages).

3.3.2 Machine Learning and Cross-Validation

We used the random forest classifier in the scikit-learn Python library to build models from each respective feature subset [39]. We selected random forest due to its effectiveness in dealing with high dimensional feature spaces, and reducing overfitting [27, 9]. It is also able to deal with highly correlated features, and provides feature importance measurements which we analyzed to find the features that were most predictive of collaboration. We cross-validated models via leave-one-group-out (each of the 20 groups used as the testing set once), and tuned hyperparameters using nested cross-validation and grid search within training data only. Hyperparameters consisted of the proportion of features to consider for each tree branch (0.25, 0.5, 0.75, or 1.0) and the minimum number of instances required in a tree node to create new branches (2, 4, 8, or 16).

Table 4 presents the values of r_{pb}, kappa, and area under the receiver operating characteristic curve (AUC) of the models, cross-validated over all data ignoring the five collaboration phases. We decided to use the point biserial correlation coefficient, r_{pb} of the true and predicted values as the primary accuracy metric, since the extreme base rates of ON-TASK and TA-CLASS behaviors (and the changing base rates of other behaviors over collaboration phases) led to unwanted sensitivity to the threshold for kappa calculation. Kappa scores (without threshold tuning) were not necessarily representative of accuracy changes as much as poorly-chosen decision thresholds. Table 4 shows that the pattern of AUC values across behavior labels was similar to r_{pb}; however, r_{pb} allows straightforward computation of confidence intervals, enabling

Table 2: Top ten most important features for each of the six considered collaborative behaviors from the combined features (onset + concurrent) random forest model. Common features in all six behaviors are in bold.

ON-TASK	ON-TASK-NO-INTERACTION	PEER INTERACTION
1. maximum seconds of no actions	1. number of actions	**1. cumul. number of page changes**
2. cumulative ratio of 2nd most to most active	**2. cumul. number of actions**	2. cumul. distance drawn
3. cumul. number of page changes	3. cumul. distance drawn	**3. cumul. number of actions**
4. number of actions	4. maximum seconds of no actions	4. cumul. ratio of 2nd most to most active
5. ratio of least to most active student	5. cumul. mean distance of same student edits	**5. cumul. number of scroll position changes**
6. proportion of students acting	**6. cumul. number of scroll position changes**	6. cumul. mean distance of same student edits
7. cumul. number of scroll position changes	7. cumul. ratio of least to most active student	7. cumul. number of tool changes
8. number of unique pages viewed	8. cumul. standard deviation of distance scrolled	8. cumul. ratio of least to most active student
9. max proportion of students on different pages	**9. cumul. number of page changes**	9. cumul. number of add object
10. cumul. number of actions	10. cumul. ratio of 2nd most to most active	10. cumul. mean y-axis value of edits
SILENT	**TASK-TALK**	**TA-CLASS**
1. cumul. number of actions	**1. cumul. number of page changes**	**1. cumul. number of actions**
2. cumul. number of selection changes	2. cumul. number of selection changes	2. cumul. standard deviation of distance scrolled
3. cumul. number of add object	3. cumul. distance drawn	**3. cumul. number of scroll position changes**
4. cumul. distance drawn	**4. cumul. number of actions**	4. cumul. maximum seconds of no actions
5. cumul. number of page changes	5. cumul. number of add object	5. cumul. proportion of students scrolling
6. cumul. number of tool changes	6. cumul. number of tool changes	**6. cumul. number of page changes**
7. cumul. mean distance of consecutive edits	**7. cumul. number of scroll position changes**	7. cumul. distance drawn
8. cumul. ratio of 2nd most to most active	8. cumul. mean distance of same student edits	8. cumul. number of add object
9. maximum seconds of no actions	9. maximum seconds of no actions	9. cumul. distance scrolled
10. cumul. number of scroll position changes	10. cumul. ratio of 2nd most to most active	10. cumul. number of selection changes

the statistical comparisons of models that we include in this paper. We thus proceeded with r_{pb} as the primary accuracy metric.

4. RESULTS

Within a 50-minute class period, we surmised that the five stages of a collaborative problem-solving team could be approximated through five equal 10-minute segments. However, if base rates of each behavior vary over time, model accuracy could as well [29]. Thus, before answering our research questions, we visualized the base rates of each behavior to help inform the results.

4.1 Base Rates Over Time

The trajectories of average base rates of the collaborative learning behaviors across these five segments of class are shown in Figure 4. Across behaviors, we observed a common pattern: the largest changes in base rates were from the first 10-minute segment to the second. The magnitude and direction of the changes in base rates during this transition were variable between the different behaviors, though some patterns can be assumed to be closely correlated. For example, the behaviors TA-CLASS and SILENT followed a similar negative trend in magnitude, since students across groups are more likely to be silent when the TA is addressing the entire class at the start of the class period, when task objectives or announcements are likely to be made. Similarly, ON-TASK, PEER-INTERACTION, and TASK-TALK tended to increase during class periods, since ON-TASK behavior is likely to involve more instances of PEER-INTERACTION and TASK-TALK behavior. ON-TASK-NO-INTERACTION showed a comparatively consistent base rate throughout the class period, perhaps being influenced by other behaviors in both directions with similar magnitude.

As base rates become more imbalanced (closer to 0 or 1),

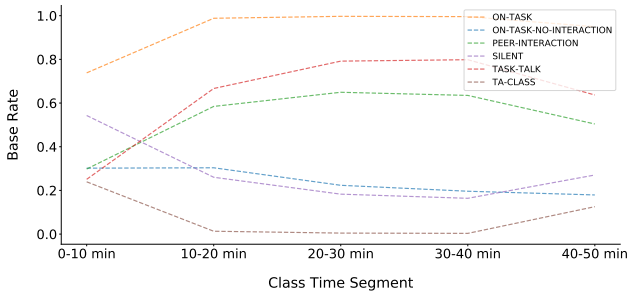

Figure 4: Average base rates of annotated behaviors across segments of each class period (averaged across class periods).

classification problems tend to become more difficult because fewer data points are available from one category of the data, and because accuracy metrics tend to become less effective [29]. Hence, the patterns in Figure 4 are important to consider when interpreting the results of the research questions.

4.2 RQ1: How does the accuracy of predicting collaborative behaviors vary across periods of a class?

To address this research question, we focused on the accuracy of the concurrent features model. This model has similar accuracy to the combined features model (see RQ3), and is more parsimonious since it has 89 features, compared to 110 features from the combined model. Thus, it will likely be the model of choice to drive predictions in future versions of the TA-facing tablet tool, and we focus RQ1 on this model.

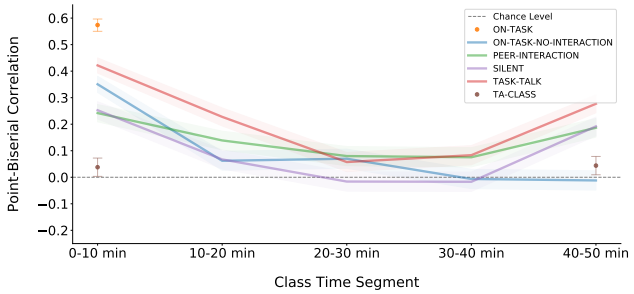

Figure 5: Concurrent features model accuracy shown over time throughout five segments of classes. Accuracy in this case consists of point-biserial correlation coefficients with 95% confidence intervals indicated by shading or by error bars for labels where base rates were too imbalanced (0% or 100%) to allow prediction in every class segment.

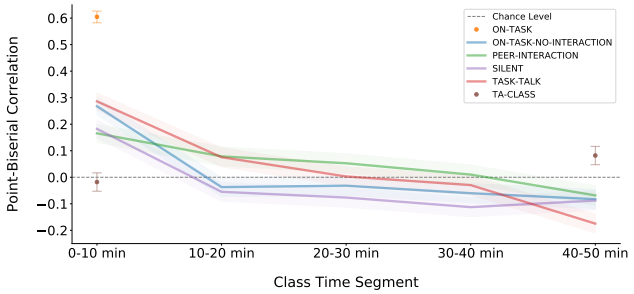

Figure 6: Point-biserial correlation coefficients with 95% confidence intervals by class segment for the onset features model.

From the overview of the model performance in Figure 5, a general U-shape pattern can be observed across the class period, where the second peak in accuracy toward the end of class never quite reached the initial accuracy from the first ten-minute segment. This trend differed for ON-TASK-NO-INTERACTION, which showed a rapid drop after the first 10-minute segment, from 0.351 to 0.062, and did not later increase. Predictions for ON-TASK-NO-INTERACTION and SILENT also briefly dropped below chance level during the second half of the class period.

4.3 RQ2: Can early class collaborative behaviors alone be used to effectively model and predict collaborative behaviors of the entire class period?

As shown in Figure 6, the onset features model had overall lower accuracy across the class periods compared to the concurrent features model (Figure 5). The absence of the U-shaped pattern from the concurrent features model (Figure 5) suggests that the first 10 minutes of collaborative behaviors may have been sufficiently similar to be captured by a model with features created from behaviors from the entire class duration, but that those behaviors were not the same as the last 10 minutes. With the exception of SILENT, predicted behaviors showed a trend toward the lowest accuracy at the end of class. However, PEER-INT remained significantly above chance for the first 30 minutes of class, indicating that groups' verbal interactions were—to a certain extent—characterized throughout most of the class period by their first 10 minutes of logged behaviors. When compared to the accuracy pattern for the concurrent model, (Figure 5), accuracy dropped below chance level more often, with ON-TASK-NO-INTERACTION, TASK-TALK, and SILENT behaviors predicted at below chance level for the latter half of the class period.

4.4 RQ3: Are collaborative behavior prediction models improved through the addition of early class collaborative behavior features, leading to a greater emphasis on the early class period?

The lower and upper 95% confidence interval bounds of the point-biserial correlation values of the models are presented in Table 3. Among the confidence intervals there was overlap for concurrent vs. combined models, but not for onset vs. concurrent and onset vs. combined (with the exception of TA-CLASS for onset vs. concurrent, not drastically so), highlighting that there was no clear significant difference in the models from the addition of onset features to the concurrent features model. This is further supported in Figure 7, which shows that the trajectory of the combined model accuracy closely resembles the concurrent features model (Figure 5). Table 4 also shows that in most cases the combined feature set was not notably better than concurrent features alone when considering overall accuracy across class time segments, in terms of r_pb, kappa, or AUC. Feature importance were analyzed and are presented in Table 2. Three common features were found in all six behaviors: *cumulative number of set page*, *cumulative number of scroll position changes*, and *cumulative number of rows*.

5. DISCUSSION

We analyzed automatic detection of collaborative problem solving in classrooms through a lens informed by small group

Table 3: Comparison of the 95% confidence intervals of the models' point-biserial correlation coefficients, r_{pb}

	Onset	Concurrent	Combined
ON-TASK	.446, .492	.506, .553	.522, .569
ON-TASK-NO-INT	.009, .040	.125, .155	.117, .146
PEER-INT	.075, .104	.247, .276	.220, .250
SILENT	.080, .112	.264, .295	.269, .300
TASK-TALK	.091, .119	.410, .438	.393, .421
TA-CLASS	-.005, .022	.012, .036	.022, .050

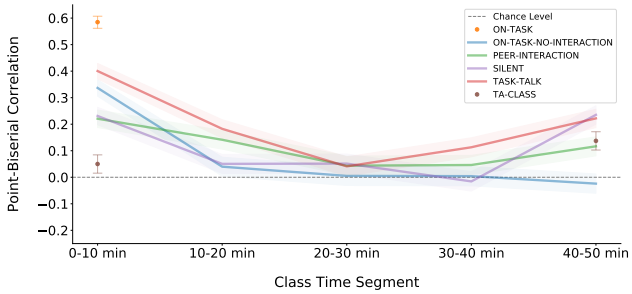

Figure 7: Point-biserial correlation coefficients with 95% confidence intervals by class segment for the combined features model.

development theory. Based on the five stages of small group development and the importance of the early periods of collaboration, we were curious if explicitly considering early class behaviors was beneficial for predicting a team's collaborative behaviors over a class period. We annotated collaborative behaviors from 16,270 video clips of an undergraduate engineering course as ground truth for behaviors, and compared the accuracy of a machine learning model built from onset features (early collaborative behaviors calculated from the first ten minutes of class) to a model from concurrent features (general collaborative behaviors over the whole class period). In this section, we discuss the implications of our findings.

5.1 Collaboration Across Class Periods

We investigated whether evidence of the five stages of group development could be seen in the concurrent model accuracy when examined in 10-minute periods of a 50-minute class. Our experimental results showed a U-shaped accuracy curve for a majority of the considered collaborative behaviors, with lowest accuracy in the middle 30 minutes of class. The base rate trends (Figure 4) may be one explanation for the observed pattern, because a majority of the behaviors also had U-shaped or inverse U-shaped base rate patterns, which is indicative of unbalanced classes. An exception to the U-shape for accuracy and base rates was ON-TASK-NO-INTERACTION, which had the highest accuracy at the beginning of class and lowest by the end. One possible explanation for this may be that the students in the first 10 minutes of the class were reading or individually thinking about the task, and transitioning to become more verbal and interactive as the class goes on—which can be approximately observed in the overall base rate pattern.

In terms of the small group development theories, the U-shape may be interpreted as evidence for the existence of three, as opposed to five, distinct stages: a beginning, a longer middle, and an end. Three stages is in line with Spitz and Sadock's three-stage model from observing the training of nursing students [47]. According to the model, stage one is characterized by anxiety-related emotions, such as curiosity and confusion, stage two is a period of trust and cohesiveness, and stage three is disengagement and anxiety about the group conclusion.

It is difficult to determine whether these stages were captured in our analysis, however, since there were some notable differences in context and aim between our study and Spitz and Sadock's research. In our study, we did not set out to capture or identify emotions during collaboration, since the focus in data collection was on annotating collaborative behaviors and capturing action data, such as tool use, scrolling, and editing. Furthermore, while previous work has developed approaches for detecting student affect through applying computer vision techniques to detect facial expressions and bodily movements on video [15, 53, 6, 7], our study used video data as means to obtain ground truth data for collaboration rather than emotion. A central goal of our research is to enable analysis for real-time collaboration intervention in the classroom, and thus we analyzed ways to detect collaboration using solely action log data, which can be applied in large and varied classroom environments even when sensors are not available. Current methods for accurately capturing emotion during learning largely rely on video or multimodal methods [5, 20], and it is difficult to envision classrooms with access to multimodal instruments and camera systems designed for analyzing emotion and collaboration.

Table 4: Accuracy comparison of the onset features, concurrent features, and combined features (onset + concurrent) models.

Behavior	Model	r_{pb}	Kappa	AUC
ON-TASK	Onset	.470	.469	.737
	Concurrent	.532	.529	.746
	Combined	.547	.545	.754
ON-TASK-NO-INT	Onset	.025	.025	.512
	Concurrent	.192	.140	.551
	Combined	.175	.131	.549
PEER-INT	Onset	.093	.090	.545
	Concurrent	.266	.261	.630
	Combined	.239	.235	.617
SILENT	Onset	.096	.096	.549
	Concurrent	.306	.279	.620
	Combined	.307	.284	.623
TASK-TALK	Onset	.115	.105	.558
	Concurrent	.445	.424	.699
	Combined	.422	.407	.692
TA-CLASS	Onset	.011	.008	.503
	Concurrent	.064	.024	.507
	Combined	.095	.036	.510

5.2 Role of Early Collaborative Behaviors

Our hypothesis that early class behaviors could effectively predict the quality of collaboration later in class was not supported by our findings. While we created onset features with characteristics of task identification of problem solving, such as verbal communication, deliberation, and reading, through features such as handwriting on the tablet, pauses between edits, high number of object removals, frequent page switches, and problem diagramming, the accuracy of the onset features model was lower than the concurrent model as a whole. Moreover, the U-shaped pattern from concurrent

features model was not observed. Despite the accuracy of the onset model showing a similarly steep decrease after the first ten minutes, it did not increase at the end of the class for any of the considered collaborative behaviors, as had the concurrent model. Taken together, the U-shape of concurrent model accuracy and the steep decline in accuracy of the onset model suggest that the first ten and last ten minutes of class are similar, but there are differences which make it difficult to effectively characterize based on features calculated from the first ten minutes of class. The similarities of the first and last ten minutes are also supported from the trends in the base rates (Figure 4). TA-CLASS behaviors—when the instructor addresses the entire class—only tend to occur at the beginning and end of class, but the content of the announcements at the beginning of the class are different and likely influence student behavior differently. For example, students may be more likely to listen and be silent in response to the announcements made at the beginning of the class since it is immediately pertinent to the class ahead, but students may be less silent when announcements are made at the class end.

Our analysis also did not support the idea that the addition of the onset features (21 features) to the concurrent features (89 features) model might improve predictive accuracy. While the resulting combined model was created using the largest number (110) of features with an emphasis on the earlier parts of class, the accuracy did not differ significantly from the concurrent features model. Comparing the confidence intervals of the model's overall point-biserial correlation coefficient (Table 3) showed that while the accuracy of onset and concurrent features models are significantly different for a majority of the behaviors except TA-CLASS (which has especially imbalanced base rates), there is overlap between concurrent and combined models for all behaviors. This indicates that the models may not be statistically different, and are not meaningfully different. Moreover, of the 110 features in the combined (onset + concurrent) model, none of the 21 onset features were found in the top eight important features in any of the six behaviors (Table 2). Three common features were found in the top eight important features for all six behaviors: *cumulative number of set page*, *cumulative number of scroll position changes*, and *cumulative number of rows*. When interpreted together, these three features may be related to the overall activity level of the groups, which may intuitively relate to changes in collaborative behavior.

6. CONCLUSION

In this paper, we were motivated by theories in small group development to analyze how explicitly accounting for early class behaviors and collaboration evolution might help improve collaboration prediction from tool action log data. We investigated collaborative problem solving in an introductory engineering course over 12 weeks. We found that collaboration prediction in a 50-minute class period did not appear to follow a straightforward interpretation of the five-stage structure, but rather a potential three-stage structure. We found that while the first ten minutes of class are distinct from the middle and ending periods of class, onset features calculated from the first ten minutes of the class could not be used to effectively predict collaboration in the later parts of class. Concurrent features (calculated from the whole 50-minute period) performed better as a whole, and the combination of onset and concurrent features did not necessarily lead to a better predicting model. Thus, groups' collaborative behaviors later in class were not notably related to their initial collaborative behaviors.

Our study was limited in several ways. Using solely tablet action log data to examine small group development restricted us from being able to account for changes in emotion, a common aspect of small group development theory. We utilized data from action logs since we wanted our analysis to be scalable and make progress toward real-time student interventions for collaboration in the classroom via prompts delivered to TAs (Figure 2). However, to promote better understanding collaborative learning theory in general, additional approaches are needed. To this end, future work examining small group development in collaborative problem solving may benefit from incorporating work on sensor-free affect detection for student engagement [2, 30], which may help identify emotions associated with various stages of small group development such as confusion or anxiety [47]. Additionally, audio of group conversations could be recorded from their tablets and aligned with action log data to understand conversations in the context of the small group development at hand. Our study was also limited by the variability of student groups in size and membership. Some groups had as few as two students in some weeks of class, and the same groups may have had four members in other weeks. This likely influenced the amount of activity captured, in addition to inevitable changes in the communication dynamic.

Insights into to the influence of early collaborative behaviors for improving collaboration prediction may help design better interventions for helping TAs facilitate collaboration, and design software tool features to promote effective student collaboration. Deeper insights into understanding small group evolution may offer ways for future work to more accurately identify a group's current collaborative stage solely from a group's behaviors, without considering content of interactions between members. Based on the assumed stage, instructors or tools could possibly allow for personalized per-team interventions to better facilitate collaborative problem solving.

7. ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. 1441149 and 1628976. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8. REFERENCES

[1] J. Andrews-Todd, C. Forsyth, J. Steinberg, and A. Rupp. Identifying profiles of collaborative problem solvers in an online electronics environment. *International Educational Data Mining Society*, 2018.

[2] R. S. Baker, S. M. Gowda, M. Wixon, J. Kalka, A. Z. Wagner, A. Salvi, V. Aleven, G. W. Kusbit, J. Ocumpaugh, and L. Rossi. Towards sensor-free affect detection in cognitive tutor algebra. *International Educational Data Mining Society*, 2012.

[3] J. Barr and A. Gunawardena. Classroom salon: a tool for social collaboration. In *Proceedings of the 43rd*

ACM technical symposium on Computer Science Education, pages 197–202, 2012.

[4] B. Barron. Achieving coordination in collaborative problem-solving groups. *The journal of the learning sciences*, 9(4):403–436, 2000.

[5] N. Bosch and S. K. D'Mello. The affective experience of novice computer programmers. *International journal of artificial intelligence in education*, 27(1):181–206, 2017.

[6] N. Bosch, S. K. D'Mello, R. S. Baker, J. Ocumpaugh, V. Shute, M. Ventura, L. Wang, and W. Zhao. Detecting student emotions in computer-enabled classrooms. In *IJCAI*, pages 4125–4129, 2016.

[7] N. Bosch, S. K. D'Mello, J. Ocumpaugh, R. S. Baker, and V. Shute. Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(2):1–26, 2016.

[8] L. J. Braaten and B. LJ. Development phases of encounter groups and related intensive groups. a critical review of models and a new proposal. *Interpersonal Development*, 1974.

[9] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[10] D. Brijlall. Exploring the stages of polya's problem-solving model during collaborative learning: A case of fractions. *International Journal of Educational Sciences*, 11(3):291–299, 2015.

[11] K. A. Bruffee. *Collaborative learning: Higher education, interdependence, and the authority of knowledge*. ERIC, 1999.

[12] P. S. Buffum, M. Frankosky, K. E. Boyer, E. N. Wiebe, B. W. Mott, and J. C. Lester. Mining sequences of gameplay for embedded assessment in collaborative learning. In *EDM*, pages 575–576. ERIC, 2016.

[13] J.-M. Burkhardt, F. Détienne, A.-M. Hébert, L. Perron, S. Safin, and P. Leclercq. An approach to assess the quality of collaboration in technology-mediated design situations. In *Proceedings of ECCE 2009: European Conference on Cognitive Ergonomics*, 2009.

[14] Z. Cai, B. Eagan, N. Dowell, J. Pennebaker, D. Shaffer, and A. Graesser. Epistemic network analysis and topic modeling for chat data from collaborative learning environment. In *Proceedings of the 10th international conference on educational data mining*, 2017.

[15] R. A. Calvo and S. K. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1):18–37, 2010.

[16] R. B. Caple. The sequential stages of group development. *Small Group Behavior*, 9(4):470–76, 1978.

[17] K. Cassidy. Tuckman revisited: Proposing a new model of group development for practitioners, 2007.

[18] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

[19] P. Dillenbourg. What do you mean by collaborative learning?, 1999.

[20] S. K. D'Mello, N. Bosch, and H. Chen. Multimodal-multisensor affect detection. In *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2*, pages 167–202. ACM, 2018.

[21] D. C. Dunphy. Phases, roles, and myths in self-analytic groups. *The Journal of Applied Behavioral Science*, 4(2):195–225, 1968.

[22] J. Garland, H. Jones, and R. Kolodny. A model for stages of development in social work groups. *Explorations in group work*, pages 17–71, 1965.

[23] A. A. Gokhale. Collaborative learning enhances critical thinking. *Journal of Technology Education*, 1995.

[24] R. Goldman, R. Pea, B. Barron, and S. J. Derry. *Video research in the learning sciences*. Routledge, 2014.

[25] P. Griffin and E. Care. *Assessment and teaching of 21st century skills: Methods and approach*. Springer, 2014.

[26] F. Hesse, E. Care, J. Buder, K. Sassenberg, and P. Griffin. A framework for teachable collaborative problem solving skills. In *Assessment and teaching of 21st century skills*, pages 37–56. Springer, 2015.

[27] T. K. Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[28] K. Huang, T. Bryant, and B. Schneider. Identifying collaborative learning states using unsupervised machine learning on eye-tracking, physiological and motion sensor data. In *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, volume 318, page 323. ERIC, 2019.

[29] L. A. Jeni, J. F. Cohn, and F. De la Torre. Facing imbalanced data–Recommendations for the use of performance metrics. In *Proceedings of the 5th International Conference on Affective Computing and Intelligent Interaction*, pages 245–251, Sept. 2013.

[30] Y. Jiang, N. Bosch, R. S. Baker, L. Paquette, J. Ocumpaugh, J. M. A. L. Andres, A. L. Moore, and G. Biswas. Expert feature-engineering vs. deep neural networks: which is better for sensor-free affect detection? In *International Conference on Artificial Intelligence in Education*, pages 198–211. Springer, 2018.

[31] J. Kang, D. An, L. Yan, and M. Liu. Collaborative problem-solving process in a science serious game: Exploring group action similarity trajectory. In *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*. ERIC, 2019.

[32] M. Kapur. Temporality matters: Advancing a method for analyzing problem-solving processes in a computer-supported collaborative environment. *International Journal of Computer-Supported Collaborative Learning*, 6(1):39–56, 2011.

[33] M. A. Kazemitabar, S. Bodnar, P. Hogaboam, Y. Chen, J. P. Sarmiento, S. P. Lajoie, C. Hmelo-Silver, R. Goldman, J. Wiseman, and L. Chan. Creating instructor dashboards to foster collaborative learning in on-line medical problem-based learning situations. In *International Conference on Learning and Collaboration Technologies*, pages 36–47. Springer, 2016.

[34] A. Lieberman. Collaborative research: Working with, not working on. *Educational leadership*, 43(5):28–32, 1986.

[35] E. Mercier, S. Shehab, J. Sun, and N. Capell. The development of collaborative practices in introductory engineering courses. In *Exploring the Material Conditions of Learning: Computer Supported*

Collaborative Learning (CSCL) Conference, pages 657–658, 2015.

[36] L. R. Offermann and R. K. Spiros. The science and practice of team development: Improving the link. *Academy of Management Journal*, 44(2):376–392, 2001.

[37] J. K. Olsen, V. Aleven, and N. Rummel. Predicting student performance in a collaborative learning environment. *International Educational Data Mining Society*, 2015.

[38] L. Paquette, N. Bosch, E. Mercier, J. Jung, S. Shehab, and Y. Tong. Matching data-driven models of group interactions to video analysis of collaborative problem solving on tablet computers. In *Proceedings of International Conference of the Learning Sciences, ICLS*, pages 312–319. International Conference of the Learning Sciences, June 2018.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, Nov. 2011.

[40] J. M. Reilly, M. Ravenell, and B. Schneider. Exploring collaboration using motion sensors and multi-modal learning analytics. *International Educational Data Mining Society*, 2018.

[41] J. M. Reilly and B. Schneider. Predicting the quality of collaborative problem solving through linguistic analysis of discourse. In *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, pages 149–157. ERIC, 2019.

[42] P. Reimann. Time is precious: Variable-and event-centred approaches to process analysis in cscl research. *International Journal of Computer-Supported Collaborative Learning*, 4(3):239–257, 2009.

[43] F. J. Reynolds and R. A. Reeve. Gesture in collaborative mathematics problem-solving. *The Journal of Mathematical Behavior*, 20(4):447–460, 2001.

[44] J. Roschelle and S. D. Teasley. The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning*, pages 69–97. Springer, 1995.

[45] P. H. Sins, W. R. Van Joolingen, E. R. Savelsbergh, and B. van Hout-Wolters. Motivation and performance within a collaborative computer-based modeling task: Relations between students' achievement goal orientation, self-efficacy, cognitive processing, and achievement. *Contemporary Educational Psychology*, 33(1):58–77, 2008.

[46] W. M. Smith. Observations over the lifetime of a small isolated group: Structure, danger, boredom, and vision. *Psychological reports*, 19(2):475–514, 1966.

[47] H. Spitz and B. Sadock. Psychiatric training of graduate nursing students. use of small interactional groups. *New York state journal of medicine*, 73(11):1334–1338, 1973.

[48] J. Thannhauser, S. Russell-Mayhew, and C. Scott. Measures of interprofessional education and collaboration. *Journal of interprofessional care*, 24(4):336–349, 2010.

[49] B. W. Tuckman. Developmental sequence in small groups. *Psychological bulletin*, 63(6):384, 1965.

[50] B. W. Tuckman and M. A. C. Jensen. Stages of small-group development revisited. *Group & Organization Studies*, 2(4):419–427, 1977.

[51] M. D. Weber and T. A. Karman. Student group approach to teaching using tuckman model of group development. *Advances in Physiology Education*, 261(6):S12, 1991.

[52] M. Yee-King and M. d'Inverno. Stimulating collaborative activity in online social learning environments with markov decision processes. In *EDM*, pages 652–653, 2016.

[53] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2008.

[54] L. A. Zurcher Jr. Stages of development in poverty program neighborhood action committees. *The Journal of Applied Behavioral Science*, 5(2):223–258, 1969.