Matching Data-Driven Models of Group Interactions to Video Analysis of Collaborative Problem Solving on Tablet Computers

Luc Paquette, Nigel Bosch, Emma Mercier, Jiyoon Jung, Saadeddine Shehab, Yurui Tong
University of Illinois at Urbana-Champaign
lpaq; pnb; mercier; jiyoonj; shehab2; ytong13 @illinois.edu

Abstract: Despite an increasing emphasis on the use of collaborative learning in classrooms, there is still much to be understood about how to successfully implement it. In particular, it is still unclear what the role of teachers should be during collaborative learning activities and how we can better support and guide teachers in their implementation of collaborative activities. In this study, we investigated how digital learning environments can be leveraged to support collaborative learning through data-driven models of students' collaborative interactions by matching video and log data. The models successfully detected off-task behavior (43.2% above chance-level accuracy) and task-related talk (34.5% above chance) as students solved problems using a collaborative sketching tool. Future work will investigate how these models can be used to allow instructors to intervene effectively to support collaborative learning through the use of data-driven tools which will provide them with live information about the students' behaviors.

Major issues and theoretical approaches

Collaborative problem solving is an important skill (Hesse, Care, Buder, Sassenberg, & Griffin, 2015), and its prominence in international education and assessment systems has been increasing (e.g. ABET, 2015; NRC, 2012; OECD, 2017). However, there is still much to be understood about how to successfully implement collaborative learning in classrooms (Nokes-Malach, Richey, & Gadgil, 2015), and in particular, how teachers can be most effective in supporting students' interactions (e.g. Webb et al., 2009). Kaendler and colleagues (2015) identified a key role that teachers play in monitoring and intervening when groups struggle, and prior work indicates the relevance of teacher interventions is important for successful group outcomes (e.g. Dekker & Elshout-Mohr, 2004). However, while master teachers are more likely to have developed expertise about assessing how and when to intervene, more novice teachers may struggle with this. For example, earlier work has shown that almost all interventions made by graduate teaching assistants were content focused, with very few interventions focused on supporting students' collaborative interactions (Mercier, Shehab & Kessler, under review). Thus, there is a need to explore ways to provide insight into the group processes for novice teachers, allowing them to understand more about what is going on within groups, and intervene appropriately (e.g. Alavi & Dillenbourg, 2012) .

In this paper, we present initial work towards creating a data-driven teacher tool that automatically provides such insight. Building on a project that sought to create a shared representation tool for engineering students, we developed prediction models by matching log data from the student tool to video analysis of their interaction behaviors. Our results indicate that there is potential in using logs of student actions to assess the quality of their interactions, which could be implemented in a teacher tool to augment their observations and provide insight into when and how best to intervene.

Collaborative learning

The value of collaborative learning for both learning and transfer and as a way to increase persistence and interest in STEM fields has been identified across a range of studies (e.g. Barron 2003; Gasiewski et al., 2012). However, variation in the quality of outcome has been recorded in both classroom and laboratory studies (Nokes-Malach et al., 2015), and success is most often associated with the quality of student interactions during collaborations (Kaendler, et al., 2015). Hesse and colleagues (2015) identified behaviors that are most associated with successful interactions, dividing them into social skills and cognitive skills. There is an increasing recognition that students need help developing these skills, and that merely placing students in groups is not sufficient for groups to function well (Authors, under review; Borge & White, 2016). Teachers play a key role in intervening to support groups as they develop these skills, needing to make a quick assessment as to whether students need support in relation to the social or cognitive collaborative processes, or in relation to the course content. Initial work in this area points towards productive insights being provided either by student actions—for example, by changing the color of a lamp (Alavi & Dillenbourg, 2012) or posting a tweet (Mercier, Rattray, & Lavery, 2015)—or by insight automatically provided to the teacher by the software the students are using (Martinez-Maldonado, Yacef, & Kay, 2015; Mericer, 2016). Thus, our primary question in this paper is how can we automatically detect students' collaborative interaction patterns and use them to provide insight to teachers?

Modeling student behavior from action logs

Researchers in the field of Educational Data Mining (Baker & Yacef, 2009) have studied how machine learning approaches can be used to build student models that are able to detect when students using digital learning environments are engaging in specific behaviors, or to infer the student's current state of mind. This is achieved by collecting detailed logs of students' actions within the learning environment. Those logs are then analyzed using machine learning algorithms to find relationships between the students' actions and the modeled construct. For example, a model might learn that repeatedly submitting the same answer on a homework problem is indicative of frustration.

Action logs have been used to model a variety of constructs across multiple types of digital learning environments, such as students' disengagement in intelligent tutoring systems, where models were trained to detect when students attempt to "game the system" (Baker, Corbett, Roll, & Koedinger, 2008; Paquette, de Carvahlo, Baker, & Ocumpaugh, 2014). Gaming the system is a type of off-task behavior in which students exploit a computerized tutor's functionalities to guess an answer or have the tutor provide them with the answer. Sabourin, Mott, and Lester (2013) studied how action logs can be used to model self-regulated learning behaviors in an educational game called Crystal Island. Gobert, Sao Pedro, Raziuddin, and Baker (2013) used action logs to assess whether students were showing behaviors related to the usage of science inquiry skills. In addition, action logs have been used to detect students' states of mind, such as their affective states (Baker et al., 2012; Kai et al., 2015; Paquette, et al., 2014; Pardos, Baker, San Pedro, Gowda, & Gowda, 2014) or whether they are mind-wandering (Mills & D'Mello, 2015). This type of research has been conducted in many types of learning environments including intelligent tutors, educational games, and science simulation microworlds.

C-STEPS

The study presented in this paper was conducted using *C-STEPS (Collaborative Support Tools for Engineering Problem Solving)*. This software is a web-based application whose main functionality is to provide students with a shared digital environment to support the creation of joint representations while engaged in collaborative problem solving. *C-STEPS* is presented on tablets, which are synchronized for each group, so members of a group share their work (and therefore, their problem-solving representations) with their teammates (Figure 1). As the students use *C-STEPS*, a detailed record of their actions is stored in logfiles.

In this paper, we focus on connecting students' actions within *C-STEPS* to video analysis of their collaborative interactions between group members. The specific research questions addressed in this paper are:

1. Are there associations between the types of interactions identified in the video and the patterns in the logfile data?
2. What insight into group processes do the logfile data provide that could inform teachers about the status of the group or appropriate intervention strategies?

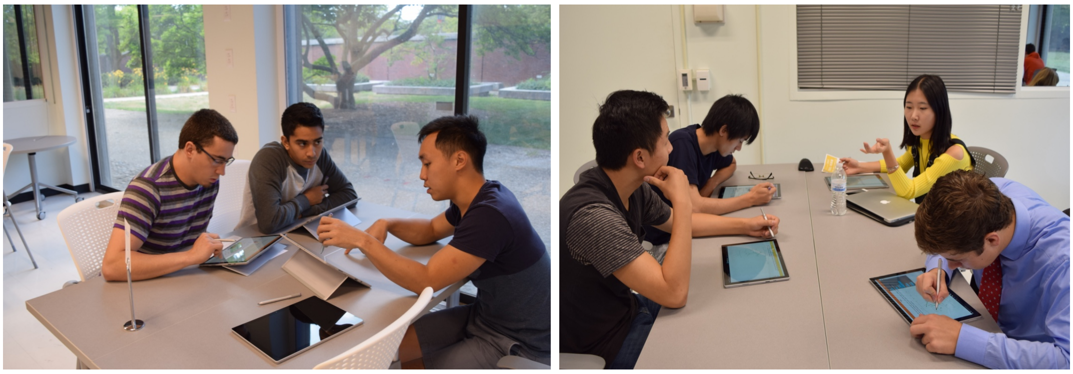

Figure 1. Students using *C-STEPS* on synchronized tablets.

Methods

Design

This study is part of a multi-year design-based implementation research project focused on collaborative learning in a large introductory engineering course. This project aligns with college goals to increase the use of collaborative learning across core introductory courses, and students engage in collaborative problem-solving activities in all discussion sections in this sequence. The research team work closely with the faculty (some who are on the research team) and TAs on task design and cultural change issues within the program. Four discussion sections (classes) attended their regularly scheduled class in an instrumented lab classroom for three consecutive weeks during which the data for this study was collected. During those sessions, students were grouped in teams of four (although due to attendance issues, group sizes ranged from 2-5) to complete collaborative engineering tasks using *C-STEPS* on either a set of tablet computers or a multi-touch table.

This paper focuses on students who used the tablet computers while solving the tasks. Multi-touch table data were not included because logged actions were fundamentally different. Data from week 1 were also discarded due to data collection issues, and as this week was seen as an introductory week for students to become familiar with the software. In total, 82 unique participants (25 female and 57 male) used the tablet computers in 14 groups in week 2 and 11 groups in week 3. All of these students gave consent to participate in the study.

Data sources

We used data collected from two sources. First, video and audio recordings were collected, providing us with rich information about how students in each group collaborated with each other. Second, action logfiles, containing a detailed list of all the students' actions on the tablets, were collected. Each logfile contained information necessary to reconstruct the students' problem-solving process. This included lists of point coordinates for each of the students' drawing segments, records of when students used clear screen and undo functionalities, screen scrolling actions, and other actions. In addition, timestamps were recorded to indicate the exact time of each action.

Data from videos and logfiles were synchronized with each other, allowing us to associate the students' actions on the tablets (logfiles) to their collaborative interactions (observed through video). The synchronized data were then segmented in one-minute clips for further analysis. We chose a length of one minute so that clips were long enough to observe meaningful collaboration behaviors while simultaneously being short enough to reduce the chance of observing multiple collaborative interactions. In total, 1,128 clips were extracted.

Coding of collaborative interaction

A subset of the videos was used to develop an emergent coding scheme of types of student interactions. The codes are shown in Table 1. The task relatedness, peer interaction, and tablet usage dimensions involved codes that were mutually exclusive; e.g., a group was either on task or off task during a one-minute clip. The talk content and teaching assistant (TA) interaction dimension were not mutually exclusive and each specific code was marked as being observed if a significant portion of the one-minute clip involved the code. For example, a clip could be marked as containing both task-related talk and other talk.

To determine reliability, two trained coders independently coded 60 video clips. The interrater reliability (Cohen's kappa) is reported in Table 1. Once interrater reliability was established, the remaining clips were coded by the two coders. Table 1 provides a count of how many instances of each code were identified in the dataset.

These codes were selected for this initial work, with a recognition of the need to expand the coding scheme to more complex collaborative behavior in later work. They are also drawn from prior research with this population, where there is evidence of TAs disrupting productive on-task collaborations when they intervene. One goal of the interface is to help TAs recognize good collaborations, as well as intervene to support groups.

Logfile feature extraction

Logfiles were processed to compute features that indicated how groups of students used the tablets during each one-minute clip. The features we computed provide information about 1) the total quantity of actions, such as how many line segments were drawn, how many times the screen was cleared, or how many time students undid their last action; 2) the location of the actions, such as the horizontal and vertical positions of drawings; and 3) student co-interaction, such as how many different students drew on the tablet during the clip and the difference between the students' scroll position. Overall, 28 features were computed from the logfile. A detailed list of those 28 features is provided in the result section (Table 3).

Detecting collaborative behaviors from logfile data

Predictive models were trained to detect the students' collaborative interactions using RapidMiner 5.3 (Mierswa, Wurst, Klinkenberg, Scholz, & Euler, 2006), a graphical data mining toolkit that allows users to apply machine learning algorithms to their data without requiring computer programming expertise. Each model was designed to detect whether students were engaged in a specific type of interaction or not. For example, whether students were off-task or not, engaging in task relevant discussion or not, and so on. Models were created for codes related to off-task behavior, no talk, task-related talk, other talk and peer interaction (Table 1). We did not create models for TA interaction, since this information would not be relevant to report to the TA, and tablet usage, since the action logs already provided us with detailed information about the groups' usage of the tablets.

Table 1. Coding scheme of student interactions, including interrater reliability and number of each code identified.

Dimension	Code	Definition and Examples	Cohen's kappa	Count
Task relatedness	On task	The clip shows that at least one of the group members appears to be on task (e.g. two students solving problems on the tablet)	0.95	905
	Off task	The clip shows that all group members appear to be off task (e.g. two students are texting using their phone)		223
Talk content	No talk	The clip shows that group members are not talking to one another or to the teaching assistant (TA)	0.88	207
	Task talk	The clip shows that at least two group members or at least one group member and the TA are talking about task-related topics		707
	Other talk	The clip shows that at least two group members or at least one group member and the TA are talking about topics that are not related to the task (e.g. socializing or technology issues)		278
Peer interaction	No peer interaction	The clip does not show any verbal interaction between group members	0.87	357
	Peer interaction	The clip shows verbal interactions between group members		771
TA interaction	Not present	The clip does not show any presence of the TA	0.91	682
	Whole class	The clip shows that the TA is interacting with the whole class (e.g. the TA is making a whole class announcement)		130
	Group	The clip shows that the TA is interacting with at least one group member		316
Tablet usage	Tablet used	The clip shows that at least one of the group members had their eyes, fingers, or stylus on the tablet	1.00	1034
	Not used	The clip shows that none of the group members were using the tablet		94

Model training

Since the type of relationship between the students' collaborative interactions and their actions on the tablet was unknown (e.g., linear or piecewise relationships), we initially tested three algorithms that capture different types of relationships: C4.5, a decision tree based approach; RIPPER, a decision rule based approach; and naïve Bayes, a probability distribution based approach. Those well-known algorithms were chosen based on related research in digital learning environments (Baker et al., 2012; Pardos et al., 2014). The naïve Bayes algorithm matched student actions best, and was selected for our final models.

Model performance evaluation

We evaluated models with two different performance indicators: Cohen's kappa (Cohen, 1960) and AUC (Area Under the receiving operating characteristic Curve) computed using the A′ approach (Hanley & McNeil, 1982). Kappa is computed in the same way as when assessing interrater agreement and measures the degree to which a model is better than chance at identifying a group's behavior. In this context, a kappa of 0 indicates a model that performs at chance level whereas a kappa of 1 indicates a perfect model. Kappa is useful for evaluating models with imbalanced codes (e.g., 80.23% on-task and 19.77% off-task behavior) because it adjusts chance level for imbalance, unlike accuracy, which may be skewed by predicting the most frequently observed interaction (e.g., 80.23% accuracy by predicting everything as on-task given its prior proportion in Table 2).

AUC was computed using the A′ approach (Hanley & McNeil, 1982). A′ is the probability that, given two examples of different codes, the model will be able to correctly classify the examples. Thus, an A′ of 0.5 indicates a model that performs at chance level, whereas an A′ of 1 indicates a perfect model. Unlike kappa, A′ evaluates the model across all possible tradeoffs between correct predictions of the positive code (the code of interest) and incorrect predictions of the negative code. This provides a complementary perspective on model performance in conjunction with kappa.

Model validation

Each model was validated using five-fold group-level cross validation. Using this approach, the full set of 1,128 clips was randomly separated into 5 subsets, each including the data for 5 of the 25 groups. Then, predictive models were trained using 4 of the 5 subsets with the remaining subset used as a held-out test set. This process was repeated five times so that each of the 5 subsets was used as the held-out test set exactly once. Performance of the model was then evaluated on the aggregated predictions of the five held-out test sets. By going through this process, we evaluated the performance of the models when predicting student interactions for new (unseen) groups of students. This was done to avoid reporting results of models that were over-fit to the training data.

Feature selection

Forward selection was used during model training, within each cross-validation fold, to find the smallest set of features that produces the best predictive model. Using this approach, features were introduced in the model one at a time, based on their predictive power. First, predictive models including only one feature were trained with each of the available features. The feature resulting in the model with the highest performance (measured as the sum of kappa and A′) was added to the set of selected features. Then, additional models were trained combining the selected feature and each of the remaining features, producing a list of models built using two features. Out of those, the best one was selected and the newly added feature was included in the set of selected features. This process was repeated, adding one new feature each time, until no increase in performance was observed.

Results

Table 2 provides a summary of the performance of the trained predictive models for each of the five predicted types of interactions. Proportions of the behaviors are reported to provide information about data imbalances. For example, in our data, 19.77% of clips were coded as off task (223 out of 1,128 clips). Note that the proportions do not sum to 100% as none of the five behaviors codes were mutually exclusive.

Table 2. Performance metrics and number of selected features for each of the five predictive models.

Type of Collaborative Interaction	Proportion	Kappa	A′	# Selected Logfile Features
Off task	19.77%	0.432	0.748	10
No talk	18.35%	0.231	0.650	8
Task talk	62.68%	0.345	0.683	7
Other talk	24.65%	0.135	0.541	14
Peer interaction	68.35%	0.225	0.682	4

We investigated which of the students' actions on the tablets were predictive of collaborative interactions by examining individual features that were selected by each predictive model. For each feature, the naïve Bayes algorithm fits two normal distributions, one for positive prediction (e.g., the students are talking about the task) and one for the negative prediction (e.g., the students are not talking about the task). This results in pairs of means and standard deviations associated to each of the logfile features in the model. We used those values to calculate effect sizes, using Cohen's *d*, that show how much the logfile features differed between predicted codes. Table 3 provides a complete list of the 28 features used in the predictive models, as well as *d* for each selected feature (blank spaces indicate that the feature was not selected).

Conclusions and Implications

Our findings show that students' action on a collaborative tool on tablet computers were indicative of their collaborative interactions with each other. As can be seen in Table 2, model performance was uneven, ranging from kappa = 0.135 to 0.432 and A′ = 0.541 to 0.748, but each model performed above chance level. We expect that the two predictive models that were most successful, off task (kappa = 0.432; A′ = 0.748) and task talk (kappa = 0.345; A′ = 0.683), will be particularly useful for informing teachers about the status of the groups as they solve the task. Indeed, although it can be easy for a teacher to observe whether students are touching their tablets; it is difficult for them to quickly evaluate whether student actions are on task without focusing their attention on each individual group. Similarly, it can be difficult for teachers to evaluate whether the students' discussions are related to the task without making an effort to listen to the content of conversations.

Further analysis of the features selected for each of our predictive models shows which of the students' actions and patterns of actions were most predictive of the types of interactions the students engaged in. As can be seen in Table 3, some features, such as the cumulative number of events for a group and the number of students who executed at least one action on their tablet within the one-minute clip, were selected for multiple models and had larger differences between behaviors (as measured using Cohen's *d*). Conversely, other features, such as the minimum and maximum horizontal positions of drawings, were infrequently selected or had small effect sizes.

Table 3. Cohen's *d* (absolute value) for each of the 28 logfile features used to build predictive models.

Logfile feature	**Off-Task**	**No-Talk**	**Task-Solving**	**Other Talk**	**Peer Inter.**
Number of total events	0.169		0.062		
Cumulative number of events in the session	0.707	0.434	0.331	0.006	
Proportion of time students spent drawing					0.532
Number of lines drawn		0.185		0.019	
Total length of lines drawn		0.207		0.323	
Number of points drawn					
Number of points erased		0.106		0.179	
Number of times students used pointer functionality (displays a temporary dot on all tablets in the group)	0.013				
Number of time entire screen was cleared (all drawing erased)	0.111				
Number of undo actions				0.202	
Number of students who drew at least once					0.704
Number of students who scrolled at least once				0.078	
Number of students who performed at least one action on their tablet	0.519	0.627	0.447		0.608
Proportion of time students spent executing actions on the tablet					
Maximum number of students simultaneously executing actions on the tablet				0.161	
Number of times students scrolled		0.007	0.025	0.127	
Standard deviation of scroll position while scrolling (captures speed)	0.100				
Range of scroll positions		0.047	0.127		
Maximum difference between different students' scroll positions	0.547		0.250	0.070	
Mean difference between different students' scroll positions				0.044	
Minimum horizontal position of drawings				0.081	
Maximum horizontal position of drawings					
Minimum vertical position of drawings			0.144	0.181	
Maximum vertical position of drawings					
Horizontal position of the center of mass of drawings	0.460	0.221		0.127	
Vertical position of the center of mass of drawings	0.445			0.032	0.309
Horizontal position of the drawing center of mass relative to the horizontal range of drawing					
Vertical position of the drawing center of mass relative to the vertical range of drawing	0.037				

As expected, the number of students who performed at least one action on their tablet during the one-minute clip, was a strong predictor of the types of collaborative interactions. This feature was selected for four of the five models and achieved some of the highest values for *d* (ranging from 0.447 to 0.627). This strong predictive power is interesting as it suggests that the students' collaboration is indeed observable in their tablet activity, rather than the collaboration only being observable outside of the collaborative tool.

The cumulative number of interaction events since the beginning of the task was also a strong predictor. It was one of only two features that was selected in four different models, and it had large *d* values (up to 0.707) in most models except for "other talk" ($d = 0.006$). This finding is interesting since the cumulative number of events does not simply capture the events within the current one-minute clip. Rather, it indicates how much students used the tablets since the beginning of the task. As such, it suggests that prior behavior is predictive of current collaboration.

Both the vertical and horizontal position of drawings on the worksheet were effective predictors of the types of student interactions. Vertical position of the drawing is an indicator of how far the students have made it through the worksheet, since students tend to work from top to bottom. Similarly, writing is usually done from left to right. As such, the horizontal position of actions on the worksheet can be used to identify productive work since unproductive drawing (e.g., doodling) is less likely to start at the left side of the worksheet. Overall, the vertical positions of the drawing interactions were more predictive than the horizontal positions, and the center of mass of the drawing was more predictive than other indicators of location (e.g., as minimum and maximum positions).

The students' scroll position were also predictive of key types of student interactions, perhaps because scroll features can be an indicator of progress on the task. Students who are scrolled further down are more likely to have made more progress towards completing the task. The strongest predictor related to scroll position was the maximum difference between students scrolling position, an indicator of whether students are paying attention to the same part of the worksheet.

This work is in its early stages and future data collection and model refinement will be necessary to improve predictions and incorporate them into tools for teachers. However, the potential of this method is clear from these initial findings. Future work will focus on improving models by taking advantage of additional data sources that are available in a live classroom setting. For example, Viswanathan and VanLehn (2017) have successfully combined audio data, collected using unidirectional headset microphones, with action logs to identify different types of collaboration. Although distributing individual headsets to each student in a live classroom is not feasible, we are investigating how the audio captured by the tablets' integrated microphones can improve our models. Similarly, future work will explore how the tablets' accelerometers can be used to improve models. Data from accelerometer could be used to give us insights about when students turn and move their tablets—for example, to show their perspective to a teammate. In addition, this early work identified simple interaction behaviors, while later work will address behaviors drawn from the research on successful groups.

Supporting students during collaborative learning is essential for effective implementation of this form of pedagogy, which is being used more extensively across higher education. However, we have found that instructors rarely have the expertise to assess and intervene successfully to support collaborative interactions (Mercier, Shehab & Kessler, under review), focusing almost exclusively on content rather than feedback to groups about their problem-solving processes. There is value of giving instructors real-time insight into group processes during class, rather than relying on their prior knowledge of collaboration to decide how to intervene or retroactively analyzing groups through video analysis. By matching tablet action logs to video analysis, we plan to leverage the automatic analysis of student actions on tablets to give instructors insight into which groups need attention, and perhaps even guidance about the types of intervention that might be needed. Such guidance will not only improve collaborative learning for students, but also teach instructors about collaborative interactions and help them to better assess groups themselves. Further research will address the question of how to best provide insights to the instructor in an actionable and meaningful way.

References

ABET. (2015). Criteria for Accrediting Engineering Programs: Effective for Reviews During the 2014-2015 Accreditation Cycle. Engineering Accreditation Commission.

Alavi, H. S., & Dillenbourg, P. (2012). An ambient awareness tool for supporting supervised collaborative problem solving. IEEE Transactions on Learning Technologies, 5(3), 264–274.

Baker, R. S. d., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. User Modeling and User-Adapted Interaction, 18(3), 287–314.

Baker, R., Gowda, S., Wixon, M., Kalka, J., Wagner, A., Salvi, A., … Rossi, L. (2012). Towards Sensor-free automated detection of affect in a Cognitive Tutor for Algebra. In Educational Data Mining 2012.

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

Barron, B. (2003). When smart groups fail. The Journal of the Learning Sciences, 12(3), 307–359.

Borge, M., & White, B. (2016). Toward the Development of Socio-Metacognitive Expertise: An Approach to Developing Collaborative Competence. *Cognition and Instruction*, *34*(4), 323–360.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46.

Dekker, R., & Elshout-Mohr, M. (2004). Teacher interventions aimed at mathematical level raising during collaborative learning. Educational Studies in Mathematics, 56(1), 39–65.

Gasiewski, J. A., Eagan, M. K., Garcia, G. A., Hurtado, S., & Chang, M. J. (2012). From gatekeeping to engagement: A multicontextual, mixed method study of student academic engagement in introductory STEM courses. Research in Higher Education, 53(2), 229–261.

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. Journal of the Learning Sciences, 22(4), 521–563.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29–36.

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In Assessment and teaching of 21st century skills (pp. 37–56). Springer.

Kaendler, C., Wiedmann, M., Rummel, N., & Spada, H. (2015). Teacher competencies for the implementation of collaborative learning in the classroom: A framework and research review. Educational Psychology Review, 27(3), 505–536.

Kai, S., Paquette, L., Baker, R. S., Bosch, N., D'Mello, S., Ocumpaugh, J., … Ventura, M. (2015). A Comparison of Video-Based and Interaction-Based Affect Detectors in Physics Playground. International Educational Data Mining Society.

Martinez-Maldonado, R., Yacef, K., & Kay, J. (2015). TSCL: A conceptual model to inform understanding of collaborative learning processes at interactive tabletops. International Journal of Human-Computer Studies, 83, 62–82.

Mercier, E. (2016). Teacher orchestration and student learning during mathematics activities in a smart classroom. International Journal of Smart Technology and Learning, 1(1), 33–52.

Mercier, E., Shehab, S. & Kessler, M. (under review) Learning to collaborate by collaborating? A longitudinal analysis of groups of engineering undergraduate students.

Mercier, E., Rattray, J., & Lavery, J. (2015). Twitter in the collaborative classroom: Micro-blogging for in-class collaborative discussions. International Journal of Social Media and Interactive Learning Environments, 3(2), 83–99.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). Yale: Rapid prototyping for complex data mining tasks. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 935–940). ACM.

Mills, C., & D'Mello, S. (2015). Toward a Real-Time (Day) Dreamcatcher: Sensor-Free Detection of Mind Wandering during Online Reading. International Educational Data Mining Society.

National Research Council. (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Washington, DC: National Academies Press.

Nokes-Malach, T. J., Richey, J. E., & Gadgil, S. (2015). When is it better to learn together? Insights from research on collaborative learning. Educational Psychology Review, 27(4), 645–656.

OECD (2017) PISA collaborative problem solving framework. Retreived from https://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

Paquette, L., Baker, R. S., Sao Pedro, M. A., Gobert, J. D., Rossi, L., Nakama, A., & Kauffman-Rogoff, Z. (2014). Sensor-free affect detection for a simulation-based science inquiry learning environment. In International Conference on Intelligent Tutoring Systems (pp. 1–10). Springer.

Paquette, L., de Carvahlo, A., Baker, R., & Ocumpaugh, J. (2014). Reengineering the Feature Distillation Process: A case study in detection of Gaming the System. In Educational Data Mining 2014.

Pardos, Z. A., Baker, R. S., San Pedro, M., Gowda, S. M., & Gowda, S. M. (2014). Affective states and state tests: investigating how affect and engagement during the school year predict end-of-year learning outcomes. Journal of Learning Analytics, 1(1), 107–128.

Sabourin, J., Mott, B., & Lester, J. (2013). Utilizing dynamic bayes nets to improve early prediction models of self-regulated learning. In International Conference on User Modeling, Adaptation, and Personalization (pp. 228–241). Springer.

Viswanathan, S. A., & Vanlehn, K. (2017). High Accuracy Detection of Collaboration From Log Data and Superficial Speech Features. Philadelphia, PA: International Society of the Learning Sciences.

Webb, N. M., Franke, M. L., De, T., Chan, A. G., Freund, D., Shein, P., & Melkonian, D. K. (2009). 'Explain to your partner': teachers' instructional practices and students' dialogue in small groups. Cambridge Journal of Education, 39(1), 49–70.

Acknowledgments

This work was supported by the National Science Foundation (NSF) through the NSF Cyberlearning and Future Learning Technologies Program (award numbers: 1441149 and 1628976).