Automated Collaboration Assessment Using Behavioral Analytics

Nonye M. Alozie, Svati Dhamija, Elizabeth McBride, and Amir Tamrakar
Maggie.alozie@sri.com, svati.dhamija@sri.com, beth.mcbride@sri.com, amir.tamrakar@sri.com
SRI International

Abstract: The 21st century skills and STEM learning standards include collaboration as a necessary learning skill in K-12 science education. To support the development of collaboration skills among students, it is important to assess and support students' proficiency in collaboration. We present the process of developing a tool that assesses collaboration quality based on behavioral communication at individual and group levels. The assessment tool uses behavior analytics comprised of multistage machine learning models built on an intricate collaboration conceptual model and coding scheme. Our collaboration conceptual model shows how layers of behavioral cues contribute to collaboration and serves as the foundation of an automated assessment tool for collaboration. We present initial findings that show reliability between our assessment of behavioral interactions with and without speech. A future automated collaboration assessment tool will give teachers information about student collaboration and help inform instruction that will guide and support students' collaboration skill development.

Key words: Collaboration, machine learning, automated assessments, science learning and instruction

Objective

The world is becoming more digital and the current professional landscape increasingly requires competency in what were once considered "soft skills," like collaboration and communication (Van Laar, Van Deursen, Van Dijk, & De Haan, 2017). The Next Generation Science Standards (NGSS Lead States, 2013) and Common Core State Standards (CCSS Initiative, 2010) each call out collaboration as a practice required for successfully engaging in STEM fields. In many K-12 classrooms, teachers use instructional methods that utilize collaboration, such as project-based learning (Krajcik & Blumenfeld, 2006) or problem-based learning (Davidson & Major, 2014) to facilitate developing student proficiencies in science and math. On a high level, collaboration is a process through which a group of people constructively explore their ideas to search for solutions that extend beyond the limited vision of one individual (Graesser, Fiore, Greiff, Andrews-Todd, Foltz, & Hesse, 2018). Incorporating collaboration during instruction can lead to increased student learning (Sung, & Hwang, 2013), since it requires group members to converge on thought processes, goals, and behaviors. Collaboration has been shown to be an effective way to achieve higher levels of understanding and robust outcomes, and is a highly sought-after professional skill (Van Laar, Van Deursen, Van Dijk, & De Haan, 2017). Since studies show the benefits of integrating collaboration into instruction and the importance of collaboration as a critical career building skill, teachers need to be able to support students' development and mastery of this skill. Teachers may incorporate the use of collaboration rubrics or peer surveys to gain additional insight about the quality of the collaboration during group activities, but due to a wide range of behavioral cues, it may be hard to determine specific behaviors that contribute to or detract from the collaboration process (e.g., Taggar & Brown, 2001; Loughry, Ohland & Moore, 2007).

Recent advances in technology, combined with a deep understanding of productive collaboration, have allowed us to begin development of a breakthrough technology that can help teachers identify key nonverbal collaborative behaviors and assess overall collaboration quality. The tool uses behavior detection to describe how well students are collaborating at individual and group levels. This is especially relevant to working in a classroom setting, since teachers have many students and often need to assess their students' collaboration from across the room or without being able to hear students clearly. In addition, focusing on behavioral cues will allow collaboration quality to be assessed across content domains, since speech is not the focus of analysis. By training the tool using multi-stage predictive machine learning models based on video analytics we will be able to automatically detect and report on the overall quality of collaboration as well as on specific behaviors that students exhibit. The tool will then be able to give teachers information about student collaboration to help inform instruction that will guide and support students' collaboration skill development. Our work provides descriptions and empirical evidence of designs that support learning that can be applied to various learning spaces and transferred to different content domains. Our design work is based on a vision of providing efficient ways to assess collaboration and provide students feedback based on their behavioral contributions to collaborative activities.

The goal of our work is to determine whether behavioral cues can accurately assess collaboration among middle school students. Currently, this work is not focused on student achievement or the outcomes of the collaboration, but instead on the *process* of collaboration: the behavioral interactions students engage in while they work collaboratively in groups. This work is in its early stages of development and we have chosen to first focus on understanding how to assess student behavior when working in collaborative groups. Our current work seeks to answer the following research question: Can visual behaviors alone be used to assess collaboration skills and collaboration quality?

Significance

This study is designed to increase knowledge about how to use automated tools to assess complex interactions that lead to collaboration. The eventual video-based analytics and machine learning can change how we collect data on classroom interactions and the future landscape of research in group learning environments, providing a new understanding of collaboration and other interpersonal interactions in learning spaces. In the future, this tool can be used as a teacher support tool; providing teachers with information on students' behavior in group settings during collaborative activities and inform future instruction that supports student ability to collaborate.

Methodology

Defining Collaboration

In this work, we developed a collaboration conceptual model (CCM; see Figure 1) that shows how various behaviors work together at different levels to promote collaboration. Our research-informed domain-independednt collaboration conceptual model delineates and reconstructs different layers of interactions that make-up collaboration. We then created a collaboration rubric for human annotation based on the CCM that assigns non-verbal behaviors to individual and group contributions to collaboration at each level, as well as the overall quality of collaboration.

Collaboration is described as a process through which a groups of people constructively explore their ideas to search for solutions that extend beyond the limited vision of one individual (Graesser, Fiore, Greiff, Andrews-Todd, Foltz, & Hesse, 2018). To assess collaboration, we needed a precise definition of collaboration or the ability to identify components of collaboration. However, collaboration is hard to define because there is complexity underlying the many seemingly simple definitions. To address this complexity, our CCM consists of tiers, or levels, of collaboration to illustrate how simple behaviors aggregate and combine into complex interactions. Research on using machine learning and behavior analytics that identify and assess group behaviors has helped us narrow our definition of collaboration and determine constructs in the development of our CCM that organizes those behaviors into individual and group interactions. The CCM is based on theoretical models that integrate research on social factors (i.e., group perceptions and personalities), cognitive science (i.e., social-cognitive systems), and education research (i.e., problem solving strategies) that capture the iterative nature of collaborative interactions.

We developed the CCM using studies that worked to parse out the complexity of collaboration through the use of constructs like teamwork and cooperative learning. Teamwork refers to the structural and interpersonal interactions between team members. Tambe (1997) devised a model for teamwork called a Shell for TEAMwork (STEAM), built using the SharedPlans Theory (Grosz, 1996; Grosz & Kraus, 1996) and Joint Intentions theory (Lavesque, Cohen, & Nunes, 1990) to operationalize a set of domain independent rules that describe how teams should work together. We used models like STEAM to build Level A of the CCM, where overall group collaboration is measured using observations of member participation and labor distribution. Cooperative learning requires students to work in small groups to achieve a shared set of goals (Johnson & Johnson, 2008). Cooperative learning focuses more on individualized behaviors and interactions among group members, like exchanging resources and information and explaining or elaborating information. Johnson and Johnson (2005) base their work on Social Interdependence Theory (Johnson & Johnson, 2008) where elements like promotive interaction and individual accountability make up cooperative learning. Cooperative learning played an instrumental role in defining Levels B1 and B2 of the CCM. We also integrated research in behavioral analytics examining the roles of affect and emotion (e.g., frustration and boredom) in collaboration when defining levels B1 and B2. Level B1 provides a description of the group dynamic that is generated among the group members and Level B2 identifies roles that each group member plays during collaborative interactions.

While the actions that make up high quality collaboration are complex, there are fine-grained nonverbal behavioral markers associated with productive collaboration (Bamaeeroo & Shokrpour, 2017). For example, Godwin, Almeda, Petroccia, Baker, and Fisher (2013) showed that on-task behavior is characterized by children directing their eye gaze at the teacher, the instructional activity, or toward appropriate instructional materials, and

that off-task behavior was when a child was looking elsewhere. We used research on individual behavioral actions that contribute to or detract from collaboration to serve as Levels C and D of the CCM. Level C delineates various complex group behaviors and Level D names the low-level features of each group member in order to explain how individual behaviors aggregate into complex interactions (Tamrakar et. al., 2012).

Collaboration cannot be effectively studied without addressing differences in how people interact due to culture, learning abilities, ethnicities, and gender. Jones (2010) warned of the skewed effects created by assuming that one data sample is as good as the next and emphasized that cultures differ in fundamental ways. Due to potential cultural and geographic differences in how individual and group behavioral interactions may be interpreted, we conducted a search of various research sources that describe effective collaborative behaviors and included behaviors that were common across sources. For example, the presence of social anxiety can vary by culture (Heinrichs, Rapee, Alde, Bogels, Hofmann, Oh, & Sakano, 2006) and can therefore interfere with how people interact with one another in group settings. A study by Kim, Yang, Atkinson, Wolfe, and Hong (2001) points to differences in socialization practices between boys and girls in various Asian cultures, which can contribute to how boys and girls participate in group settings. Even with various differences among collaborators, studies show that diversity of behaviors produces higher quality outcomes (e.g., Barjak & Robinson, 2008). The development of the CCM incorporates these findings by surveying students about their collaboration comfort levels, adjusting the rubric to address how students self-identified, and include literature-based behaviors that have been validated across various populations.

We continue to refine our collaboration conceptual model and rubric through piloting and continued literature reviews to accommodate diverse behavioral norms. This CCM is used with behavior-based learning analytics to train machine learning models to assess the quality of collaboration among small groups of students in face-to-face collaboration settings. We analyzed collaboration using individual student behaviors (Levels C and D), as well as overall collaboration and participation structures (Levels A and B). Our unit of analysis is at both the individual and group levels, since individual behaviors impact group behavior and overall collaboration.

Stratifying the Conceptual Model for Machine Learning

Figure 1 shows how our Multimodal Integrated Behavior Analysis (MIBA) software extracts low level tracking of human head-pose, eye gaze, facial expressions, body-pose and gestures in Level E. The low level features from Level E are used to generate Level D descriptors like joint attention and smiling. The Level D descriptors are used to describe more complex interactions, such as "sharing tools" or "explaining ideas", in Level C. Complex behaviors from Level C are used to determine the individual roles of each student, such as "follower" or "group guide" in Level B2, and group dynamics like "social and hyperactive" in Level B1. All levels come together as an overall collaboration code, such as "effective" or "progressing," in Level A.

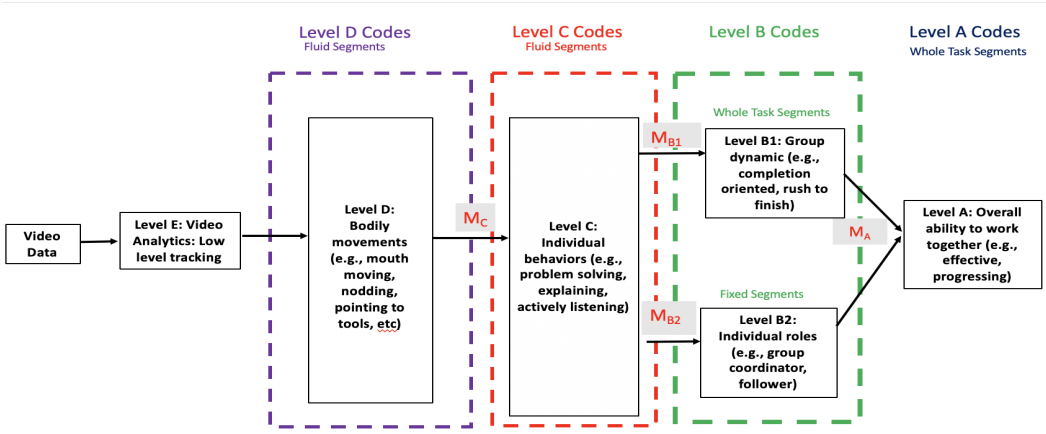

Figure 1. The collaboration conceptual model.

Data Collection

In the spring of 2019, we collected 15 hours of video data in five middle schools. Sixty volunteer students completed a brief survey that collected information about their demographics (e.g., languages spoken, ethnicity) and comfort levels with collaboration and various science concepts. We videotaped and audio recorded one group at a time. Students worked in small groups for one hour to complete up to 12 open-ended life science and physical

science tasks that required students to construct models of science phenomena. We used short, structured tasks to localize student behaviors. We collected data by positioning three Microsoft kinect cameras in a triangular configuration around groups of four students, providing full-body 3D motion capture, facial recognition and voice recognition (see Figure 2). Students were given logistic and organizational instructions, but did not receive help during the completion of the task. Students were instructed to work together for one hour or until the tasks were completed.

We administered 2 pre-pilots in preparation for the pilot study. During pre-pilot 1.0, we collected 1 hour of video data of students working on science tasks in a collaborative group. The tasks assigned in pre-pilot 1.0 were further refined and tested in pre-pilot 2.0. During pre-pilot 2.0, we collected another hour of video data and tested the upgraded equipment, data collection instruments, and analysis techniques before collecting data for the pilot study. We used the video from both pre-pilots to train the human annotators.

Figure 2. The data collection setup.

Science Task Design

Students completed up to 12 open-ended tasks in photosynthesis, cellular respiration, energy and transfer, and ecosystems that required them to develop models or solve a scientific problem. Half of the tasks asked students to arrange physical manipulatives (pieces of paper with images, words, or graphs on them) in addition to developing an explanation. For example, in one modeling task, students arranged physical manipulatives that depicted a banana peel, an orange peel, soil, bacteria, fungi, and rich soil nutrients in order to develop a model for how decomposition happens (see Figure 3). In tasks without physical manipulatives, students analyzed graphs or data tables, developed explanations, or drew their own models for different types of systems.

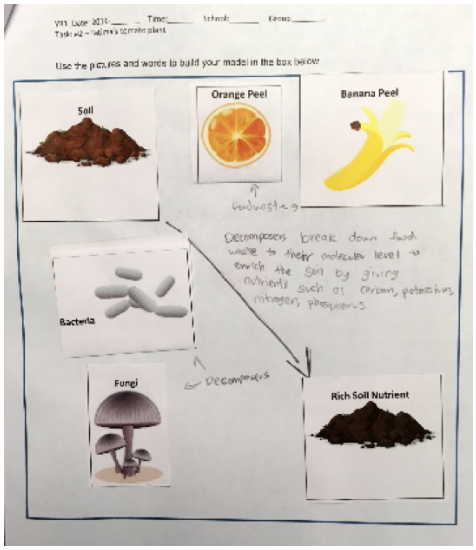

Figure 3. The figure above shows a sample response to a modelling task.

Data Annotation

Each video recording was coded by human annotators using ELAN (an open-source annotation software). The video recordings were segmented by task with a maximum of 12 segments per video (1 segment per task). Collaboration tasks ranged from 5 to 25 minutes each. Video recordings were also separated into two modalities, (1) *video modality* (assessing visual only), and (2) *full modality* (assessing visual and sound). Each video was first manually coded at each level (levels shown in Figure 1) using our collaboration assessment rubric, by three different education researchers in the video modality. The full modality videos were coded after the video modality to prevent bias. The human annotators were mixed so that each video was coded by 3 different researchers each time. The majority vote from the group of three coders was used as the gold standard for the annotation.

Analysis & Findings

Analysis 1: IRR for visual behaviors of collaboration

To establish interrater reliability (IRR) and for code refinement, the human annotators participated in multiple training sessions over approximately 3 months using data from 2 pre-pilot sessions for coding the A and B levels. The video recordings from the pre-pilot sessions were used to train the human annotators and the annotators were not allowed to access the pilot video-recordings during that time. During training, human annotators were assigned a segment of video from the pre-pilot data to code individually, then the annotators discussed their codes in a group. Through discussions about IRR, each code was refined to increase accuracy and agreement in behavior identification. At the end of the training, the A and B level codes were evaluated for rater agreement (i.e. majority votes) and a Cohen's Kappa score. Cohen's Kappa IRR for impoverished coding of levels A and B was calculated to reduce effects of any agreement that could have occurred due to random chance. Table 1 shows that the rater agreement score for the training data ($RAgreement_{TR}$) was between 0.5-0.66 and a Cohen's Kappa score ($Kappa_{TR}$) of 0.27-0.4, showing fair agreement (Landis & Koch, 1977; Viera & Garrett, 2005).

After training, each pilot video was independently coded 3 times to account for annotation bias in the video modality. IRR was calculated for levels A, B1, and B2 (refer to Table 2 for details), which required the human annotators to assess collaboration at higher levels, involving the culmination of complex behaviors and a nuanced understanding of human-human interactions *without the use of sound.* The pilot data had a rater agreement range ($RAgreement_{Video}$) between 0.63-0.68 and a Cohen's Kappa range ($Kappa_{Video}$) of 0.47 - 0.49, showing an increase to moderate agreement. We also determined the Cohen's Kappa scores for interrater reliability of the full modality data to establish validity with the collaboration conceptual model and collaboration coding rubric. Using the pilot data, we were able to establish that the annotator codes were consistent with the collaboration conceptual model. Because the annotators were coding *with sound*, they achieved a higher Cohen's Kappa IRR ranging from 0.5 - 0.56. Moderate agreement for IRR calculation is comparable to other machine learning research that calculates IRR for human annotations using sound (e.g., Lubold and Pon-Barry, 2014; Richey, D'Angelo, Alozie, Bratt, & Shriberg, 2016). Periodically, codes were checked for accuracy by non-annotators against the collaboration rubric to monitor validity. This step is critical for the accuracy of the collaboration conceptual model and for comparisons between video and full modality coding.

Table 1: The inter-rater agreement results in the form of rater agreement (RA) and Cohen's Kappa scores for training data (TR) (video modality)

Collaboration skill level	$\boldsymbol{RAgreement_{TR}}$	$\boldsymbol{Kappa_{TR}}$
A-Level: Collaboration Quality	0.55	0.41
B1-Level: Group dynamics	0.66	0.27
B2-Level: Individual role/ participation	0.55	0.40

Table 2: The inter-rater agreement results in the form of rater agreement (RA) and Cohen's Kappa scores for pilot data (video and full modalities)

Collaboration skill level	$\boldsymbol{RAgreement_{Video}}$	$\boldsymbol{RAgreement_{Full}}$	$\boldsymbol{Kappa_{Video}}$	$\boldsymbol{Kappa_{Full}}$
A-Level: Collaboration Quality	0.675	0.709	0.49	0.5

B1-Level: Group dynamics	0.680	0.715	0.47	0.53
B2-Level: Individual role/ participation	0.639	0.689	0.48	0.56

Analysis 2: Data distribution

Pilot data consisted of 60 middle school students collaborating in small groups of 3 or 4. Students were recruited from schools in a suburban area surrounding a major city in the United States and were from high performing schools, showing above state average standardized test scores. Student surveys showed that 50% of all participating students perceived themselves as very comfortable with collaboration, while there was mixed comfort with the science concepts. Additionally, a significant majority of Asian male students volunteered for the study, and attended schools where the student population is over 80% Asian, thereby introducing bias to the data.

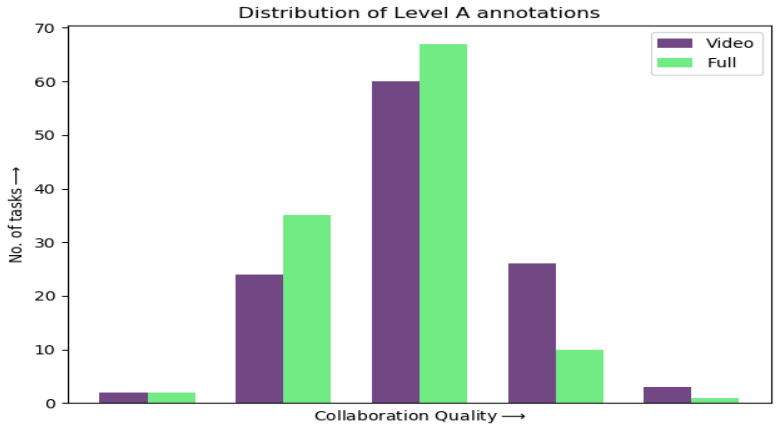

Figure 4. The figure above shows the distribution of annotations at the A-Level.

After calculating interrater agreement for video data, we analyzed the distribution of the annotations across modalities (video only vs video and sound) of the pilot data. In the pilot study, a total of 117 tasks were annotated by the education researchers. Our initial analysis was a comparison between the video and full modality annotations at the A (overall collaboration quality) and B1 (group dynamics) levels. We show a distribution of the annotations (A-Level codes shown in Figure 4) comparing video and full modality codes and found them to be comparable for most of the overall collaboration quality assessment codes. Figure 4 shows the number of tasks that were assigned specific codes for the A-Level (the overall quality of collaboration) in video modality (left bar) and full modality (right bar). We excluded 2 tasks due to lack of majority in agreement (Note: due to non-disclosure agreements, we are unable to provide the names of the codes at this time). The far left of the graph represents the lowest quality of collaboration and the far right represent the highest qualities of collaboration. Most students were coded as needing support in learning how to collaborate. Very few students were found to be effective collaborators, regardless of their willingness to participate in the study. It is worth noting that the lack of audio in the video modality had little to no impact on the assessments of lower quality group collaborations, compared with using both video and audio.

The data distribution for level B1 (group dynamics) was similar to the A-Level data distribution; annotations between video and full modality codes were comparable. At the B1-Level. As mentioned earlier, the data was biased towards high performing students who volunteered to be in the study, shifting the annotations toward codes that show the group dynamics as focused, calm, and work oriented. However, the similarity in the distribution of data does not imply high agreements at a task-level or feasibility of using either modality for automated assessment. Therefore, we further analyzed the annotations at the task level empirically in Analysis 3.

Analysis 3: How does impoverished modality compare to the full modality?

In order to develop an automated tool that could assess collaboration skills using visual behaviors, we needed to test whether visual behaviors alone could be used to estimate collaboration skills and quality. To compare annotations for the video modality with the full modality, we assigned the same set of videos to the same annotators. Annotators coded the video modality prior to coding the full modality, and a period of multiple months elapsed between the video modality coding and the full modality coding. The interrater agreement comparison for audio and video modality on a per annotator basis is shown in Table 3 below.

Table 3: The table below shows the inter-rater agreement results comparing video modality vs full (audio and video) modality

Collaboration skill level	$RAgreement_{Full\|Video}$	$Kappa_{Full\|Video}$
A-Level: Collaboration Quality	0.56	0.29
B1-Level: Group dynamics	0.63	0.40
B2-Level: Individual role/ participation	0.66	0.51

When we compared the video and full modality codes, we found Cohen's Kappa score to be fair at the A-level (i.e. 0.29) and moderate at B1 and B2 levels, ranging from 0.40 – 0.51. Possible reasons for the wide range of kappa scores are 1) A-level annotations require a higher level of contextual knowledge and inference than B1 and B2 level annotations, 2) the number of groups in our pilot data set was too small to show enough variation, and 3) the data sameples were from a select group of volunteer students and biased the annotations to a narrow selection of annotation options, increasing the possibility of agreement by chance. This can lower the kappa score significantly. Although this test shows promise in feasibility, it also shows the need for diverse data.

Conclusions and Next Steps

Our initial work with the MIBA system shows successful calibration and synchronization of high-quality information from audiovisual and sensory input. Using MIBA, the collaboration conceptual model has the ability to capture behavioral cues associated with collaboration with reliability and validity. Our initial analyses shows that there is potential for assessing the quality of collaboration using behavioral cues alone. We showed that the human annotations of video using sound was in moderate agreement with human annotations using no sound, indicating robustness and steadiness in our collaboration conceptual model and refinement in our collaboration coding rubric. The distribution of A and B1- Level codes indicates that most students in the study are progressing toward being proficient collaborators; where they exhibit behaviors that contribute to good collaboration. This preliminary finding indicates that many students would benefit from increased instructional support that would help them develop effective collaboration behaviors. We are currently in the process of continuing the reliability testing between modalities for levels C and D. As of now, our rater agreement for video data, using a single task with 3 coders each, ($RAgreement_{Video}$) is 0.85 and Cohen's Kappa ($Kappa_{Video}$) is 0.81, showing substantial agreement (when unlinked annotations are excluded from the calculation, thereby overestimating reliability). The completion of all collaboration levels will create opportunities to understand how low-level human behaviors contribute to high level interactions patterns that contribute to collaboration. Our current analysis for A, B1 and B2 codes is based on 15 hours of video data and requires more data to improve our reliability scores. Moving forward, we will collect additional data to further refine and validate our collaboration conceptual model and coding scheme, increase reliability, and make further comparisons between impoverished and full modality annotations.

Our current student sample did not represent a diverse student body. Our sample requires a diverse group of students to capture a variety of interaction patterns. The future development of the machine learning models depends on large amounts of diverse data to be applied to a wide range of students. Student diversity will help increase the validity of the machine learning models we will develop upon demonstrating the ability to use behavioral cues to assess collaboration quality.

Our immediate next steps following this study will be to test our data analysis in authentic classroom settings with students working on the science tasks designed for this project in groups assigned by their teachers, and explore the relationship between the quality of collaboration and the quality of student artifacts. Over time, we plan to perform impact studies in classrooms and other learning spaces to determine the kind of effect the tool has on group work, student artifact development, student productivity, and inclusion among group members. We will also explore instructional resources, such as dashboards, to support teachers in the use of this tool. These steps will initiate studying technology development and its role in measurement and improving teaching and learning.

References

Bambaeeroo, F., & Shokrpour, N. (2017). The impact of the teachers' non-verbal communication on success in teaching. Journal of advances in medical education & professionalism, 5(2), 51

Barjak, F., & Robinson, S. (2008). International collaboration, mobility and team diversity in the life sciences: impact on research performance. *Social Geography*, *3*(1), 23-36.

Common Core State Standards Initiative. (2010). Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects. Retrieved from http://www.corestandards.org/ELA-Literacy/

Davidson, N., & Major, C. H. (2014). Boundary crossings: Cooperative learning, collaborative learning, and problem-based learning. Journal on excellence in college teaching, 25.

Godwin, K., Almeda, V., Petroccia, M., Baker, R., & Fisher, A. (2013). Classroom activities and off-task behavior in elementary school children. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 35, No. 35).

Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the science of collaborative problem solving. Psychological Science in the Public Interest, 19(2), 59-92.

Grosz, B. J. (1996). Collaborative systems (AAAI-94 presidential address). AI magazine, 17(2), 67-67.

Grosz, B. J., & Kraus, S. (1996). Collaborative plans for complex group action. Artificial Intelligence, 86(2), 269-357.

Heinrichs, N., Rapee, R. M., Alden, L. A., Bögels, S., Hofmann, S. G., Oh, K. J., & Sakano, Y. (2006). Cultural differences in perceived social norms and social anxiety. *Behaviour research and therapy*, *44*(8), 1187-1197.

Johnson, R. T., & Johnson, D. W. (2008). Active learning: Cooperation in the classroom. The annual report of educational psychology in Japan (47), 29-30.

Johnson, D. W., & Johnson, R. T. (2005). New developments in social interdependence theory. Psychology Monographs, 131, No. 4.

Johnson, D. W., & Johnson, R. T. (2008). Social interdependence theory and cooperative learning: The teacher's role. In *The teacher's role in implementing cooperative learning in the classroom* (pp. 9-37). Springer, Boston, MA.

Jones, D. (2010, June). A WEIRD view of human nature skews psychologists' studies. Science, 328(5986), 1627

Kim, B. S., Yang, P. H., Atkinson, D. R., Wolfe, M. M., & Hong, S. (2001). Cultural value similarities and differences among Asian American ethnic groups. *Cultural Diversity and Ethnic Minority Psychology*, *7*(4), 343.

Krajcik, J. S., & Blumenfeld, P. C. (2006). Project-based learning (pp. 317-34).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Levesque, H. J., Cohen, P. R., & Nunes, J. H. (1990, July). On acting together. In Proceedings of the 8th National Conference on Artificial Intelligence (AAAI-90), Vol. 1, pp. 94-99, Boston, MA.

Loughry, Ohland & Moore, 2007

Lubold, N., & Pon-Barry, H. (2014, November). Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (pp. 5-12). ACM.

NGSS Lead States (2013). *Next Generation Science Standards: For states, by states.* Washington, DC: The National Academies Press.

Richey, C., D'Angelo, C., Alozie, N., Bratt, H., & Shriberg, E. (2016, January). The SRI Speech-Based Collaborative Learning Corpus. In INTERSPEECH (pp. 1550-1554). Bambaeeroo, F., & Shokrpour, N. (2017). The impact of the teachers' non-verbal communication on success in teaching. Journal of advances in medical education & professionalism, 5(2), 51.

Sung, H. Y., & Hwang, G. J. (2013). A collaborative game-based learning approach to improving students' learning performance in science courses. Computers & education, 63, 43-51.

Taggar, S., & Brown, T. C. (2001). Problem-solving team behaviors: Development and validation of BOS and a hierarchical factor structure. Small Group Research, 32(6), 698-726.

Tambe, M. (1997). Towards flexible teamwork. Journal of artificial intelligence research, 7, 83-124.

Tamrakar, A., Ali, S., Yu, Q., Liu, J., Javed, O., Divakaran, A., Cheng, H., & Sawhney, H. (2012, June). Evaluation of low-level features and their combinations for complex event detection in open source videos. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 3681-3688). IEEE.

Van Laar, E., Van Deursen, A. J., Van Dijk, J. A., & De Haan, J. (2017). The relation between 21st-century skills and digital skills: A systematic literature review. Computers in human behavior, 72, 577-588.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam med*, *37*(5), 360-363.