Enhancing Stealth Assessment in Collaborative Game-based Learning with Multi-task Learning

Anisha Gupta[1], Dan Carpenter[1], Wookhee Min[1], Bradford Mott[1], Krista Glazewski[2], Cindy E. Hmelo-Silver[2], James Lester[1]

[1] North Carolina State University, Raleigh NC, USA
{agupta44, dcarpen2, wmin, bwmott, lester}@ncsu.edu
[2] Indiana University, Bloomington IN, USA
{glaze, chmelosi}@indiana.edu

Abstract. Collaborative game-based learning environments offer the promise of combining the strengths of computer-supported collaborative learning and game-based learning to enable students to work collectively towards achieving problem-solving goals in engaging storyworlds. Group chat plays an important role in such environments, enabling students to communicate with team members while exploring the learning environment and collaborating on problem solving. However, students may engage in chat behavior that negatively affects learning. To help address this problem, we introduce a multidimensional stealth assessment model for jointly predicting students' out-of-domain contributions to group chat as well as their learning outcomes with multi-task learning. Results from evaluating the model indicate that multi-task learning, which simultaneously performs the multidimensional stealth assessment, utilizing predictive features extracted from in-game actions and group chat data outperforms single-task variants and suggest that multi-task learning can effectively support stealth assessment in collaborative game-based learning environments.

Keywords: Stealth Assessment, Multi-Task Learning, Computer-Supported Collaborative Learning, Collaborative Game-Based Learning, Game-Based Learning Environments.

1 Introduction

Computer-supported collaborative learning has been shown to effectively foster students' collaborative problem-solving skills with pedagogical strategies including inquiry learning and problem-based learning [1-3]. Collaborative game-based learning combines the benefits of both computer-supported collaborative learning, which focuses on social aspects of learning, and game-based learning environments that create engaging learning experiences. Collaborative game-based learning enables students to discuss topics, ask questions, and collectively brainstorm towards achieving shared goals with their team in a game environment. Unlike single-player game-based learning environments, students can play an active role in guiding their peers while also

receiving help from teammates towards mastering in-game concepts together [4]. In particular, collaborating in small groups helps students learn better in problem-based collaborative game-based learning, where students actively participate in engaging problem solving and collaboratively work towards learning and completing the game, while a teacher plays the role of a facilitator who provides guidance for each team [2,5].

Collaborative game-based learning environments are often equipped with online chat interfaces for students to communicate with each other [5-7]. Although this is a useful tool for students to exchange ideas while solving a shared problem, some students may engage in out-of-domain chat [6,8]. Out-of-domain messages sent in the chat could be a source of distraction to students and could negatively impact their shared learning experience. While some out-of-domain chat behavior may be constructive, it could also transition students to negative affective states such as frustration or perhaps lead to unproductive learning outcomes when it is associated with certain affective states such as confusion [9]. It would thus be desirable to detect participation in out-of-domain chat early in collaborative game-based learning environments and allocate the necessary pedagogical support to students with the aim of helping them regulate their behavior. To address this challenge, we introduce a stealth assessment model that dynamically predicts students' out-of-domain chat contributions as well as their learning outcomes. Stealth assessment, which is grounded in evidence-centered design [10], uses a stream of students' interaction data within a game-based learning environment to make inferences about their competencies [11]. While assessing individual students' knowledge has often been a focus of stealth assessment, it can be extended to support assessments in collaborative game-based learning by inferring conclusions regarding both learning and collaboration. In collaborative game-based learning environments, it can be beneficial to identify which students are engaging in out-of-domain chat behaviors during the collaborative learning experience. While some level of rapport building is expected in collaborative game-based learning environments, it is important to ensure that such interactions do not distract students from the primary learning objective. Our goal is not to eliminate all out-of-domain chat, but rather to provide a tool that can help teachers identify and address out-of-domain messages that are not conducive to learning. Accurate and early prediction of learning outcomes and contribution to out-of-domain messages can be especially useful for providing targeted pedagogical assistance to students at the early phase of students' interactions with the learning environment. For example, if certain members of a group are predicted to engage in out-of-domain behavior, then they may be impeding the learning process for the entire group, at which point the stealth assessment model can inform a facilitator or an automated scaffolding system to intervene in the online chat and redirect the focus of the conversation in the early phase of collaborative learning.

This work presents a multidimensional stealth assessment model that explores multi-task learning for simultaneously predicting post-test performance and out-of-domain contribution of students by dynamically analyzing their interactions with a collaborative game-based learning environment for teaching ecosystem science to middle school students. We explore the benefits of leveraging information from mes-

sages exchanged in the group chat, the representations of which are driven by a transformer-based language model, in addition to game trace logs, as input features to our multi-task learning stealth assessment model. To effectively model these multimodal features, we adopt a late fusion approach that concatenates chat data representations and in-game action representations generated using two separate variants of recurrent neural networks in our multidimensional stealth assessment model. A competitive baseline using random forest models was created to evaluate our deep learning-based stealth assessment models with respect to predictive accuracy of both post-test performance and out-of-domain contributions as well as the models' early prediction capacity for both predictive tasks.

2 Related Work

Stealth assessment models have been extensively explored in the context of game-based learning environments [11]. Stealth assessment is an application of evidence-centered design (ECD) [10], which performs inference of higher-level student competencies based on task-level evidence. ECD consists of three core components: a competency model that models students' knowledge and skills, an evidence model that relates observations of student behavior to the competency variables, and a task model that provides the students with problems suited to demonstrate their competency and gather evidence of their knowledge. Stealth assessment implements ECD in the context of game-based learning by using fine-grained action-level information to model student competencies without interrupting gameplay or disrupting engagement [12-13]. Stealth assessment has been extensively explored for unobtrusively measuring student learning outcomes such as problem solving [14], creativity [15], and computer science skills [16] in game-based learning environments. Multi-task learning [17], a technique for inducing shared representations by modeling multiple related machine learning tasks and leveraging them to improve predictive performance, has been investigated for stealth assessment with a goal to improve predictive performance of stealth assessment models. Gupta et al. used multi-task learning for predicting post-test scores and quality of written reflections in single-player game-based learning environments [18].

Collaboration has been shown to be highly associated with students' learning outcomes in collaborative learning environments [19]. For example, Sung et al. found a significant difference in learning outcomes between groups of students who collaborated in a game-based learning environment and students who played the game individually [20]. Given this relationship between learning outcomes and constructive collaboration, we present a stealth assessment model that simultaneously predicts post-test performance and out-of-domain chat contribution of students in collaborative game-based learning environments with multi-task learning, utilizing action logs and distributed representations of group chat messages.

Prior work has also explored natural language processing techniques to obtain embeddings of student-generated text for improving the predictive performance of stealth assessment models [18]. Pre-trained embeddings, which are vector representa-

tions of text such as words and phrases that are learned from large amounts of text data, have also been examined to represent chat messages exchanged by students in small group-based collaborative game-based learning environments. Carpenter et al. investigated Word2Vec, ELMo, and BERT embeddings for identifying out-of-domain messages in collaborative chats [8], reporting best predictive performance using pre-trained BERT embeddings. Park et al. showed that an LSTM-based model trained on pre-trained BERT embeddings of collaborative group chats outperformed other baseline models for disruptive talk detection [6]. In our current study, we present a model that predicts out-of-domain contribution over the entire course of a students' gameplay, based on evidence of previous actions and chat records, in addition to predicting their post-test performance after gameplay. Our work is first to investigate a stealth assessment model to evaluate students' collaboration and learning using multi-task learning for collaborative game-based learning environments.

3 Dataset

Our experiments were conducted on data collected from two studies using CRYSTAL ISLAND: ECOJOURNEYS (Figure 1), a collaborative game-based learning environment for middle school ecosystem science. For both the studies, students completed a pre-test measuring ecosystems content knowledge, and then they played the game for 6 classroom periods, followed by attempting a post-test assessment with the same questions as the pre-test. In the game students are tasked with identifying an illness that is affecting the fish population on a remote island. Students in the class are divided into teams that work together towards solving the problem in the game. A team typically consisted of four students, with a minimum of three students. Each member of the team explores the game environment to gather information scattered across the map. Students gather at a virtual whiteboard integrated in the game to discuss and organize their findings to collaboratively figure out what might be plaguing the fish on the island. Each team is also equipped with an in-game group chat interface, where students communicate with their teammates to discuss their findings and collaborate on ideas. A facilitator, who is a domain-expert research team member, regulated the chat for each group, prompting for group discussion when needed, taking a poll on ideas, and occasionally nudging the focus of the team back to the in-game problem solving if students were distracted.

In our current work, we used a dataset consisting of a total of 8,350 chat messages, of which 2,000 messages were sent by the facilitator and the remaining messages were sent by 72 consented and assented middle school students (11-12 years old, 31 females and 41 males) divided into 18 groups. Three students who did not complete the post-test assessment were excluded from the dataset, but their group chat messages were included as context for predicting out-of-domain contribution and learning outcomes of other students in the dataset. On average, each group chat consisted of 463.89 messages (SD=285.26), and an average of 92.78 messages were sent by each student across all the groups (SD=78.45).

Fig. 1. CRYSTAL ISLAND: ECOJOURNEYS collaborative game-based learning environment

3.1 Out-of-Domain Labeling

Our chat data comprises messages collected from both studies using the collaborative game-based learning environment. Chat messages from the first study were previously labeled by two researchers [21], where the inter-rater reliability measured with Cohen's kappa is 0.751, indicating substantial agreement. The chat messages from the second study were labeled by two researchers, including one rater who labeled the dataset from the first study. Cohen's kappa of 0.878 was achieved between the two raters on the chat messages from the second study, indicating almost perfect agreement. All three researchers involved in labeling the chat messages for both the studies labeled 20% of the first dataset in common and achieved Fleiss' kappa of 0.90, indicating very good agreement. A total of 2,259 out of 8,350 chat messages were labeled as out-of-domain in the combined dataset (M=132.88, SD=107.75 per group; M=28.96, SD=42.21 per student). The rubric used for labeling the chat dataset for both the studies is shown in Table 1 [21].

Table 1. Rubric for labeling chat dataset as on-task or out-of-domain.

Label	Rubric	Example
On-task	Text that contributes to the science discussion in the group, demonstrates or addresses a relevant affective state, fosters collaboration, or asks a relevant question	*"Im at the whiteboard now, what do I do?"* *"Tilapia need warm temperatures in their water, dissolved oxygen, and food (Cant remember what food)"*
Out-of-domain	Text that is meaningless or unintelligible, or off-topic conversation that is unrelated to learning outcomes and fails to address affective states of students in the group	*"REEEEEEEEE"* *"are u boy or girl"* *"WHATS A TURTLELE"* *"gonna be as mean as possible"*

3.2 Post-test Assessment

As noted above, we utilize data collected from two separate studies that used CRYSTAL ISLAND: ECOJOURNEYS. These studies had slightly different pre- and post-test assessment questionnaires. The first study had a total of 42 questions. Based on the outcome of the first study, item analysis was performed on the test content to shorten its length and the second study was conducted with 32 assessment questions that were directed towards the same learning goals. For the post-test performance prediction task, we predict performance labels driven by the normalized score for each student. The distribution of normalized post-test scores across data collected from both studies is shown in Fig. 2.

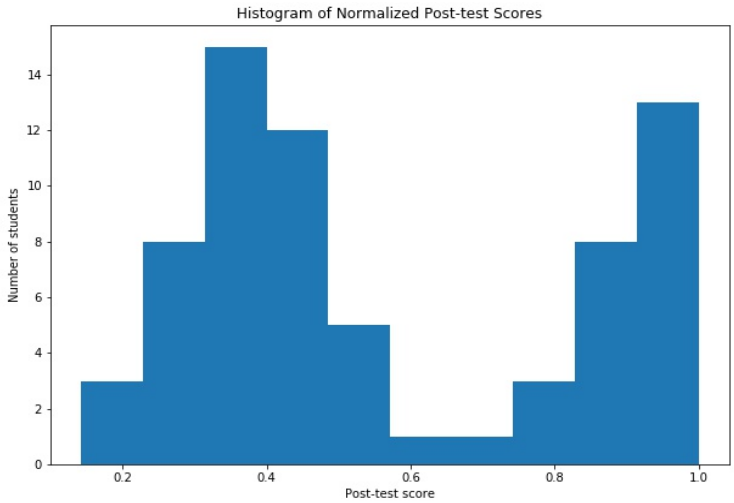

Fig. 2. Histogram of normalized post-test score distribution across both studies

The post-test questions were presented to the students in various formats such as multiple choice, interpreting charts and tables, and classification problems. Questions included problems directly based on in-game text, such as classifying components of an ecosystem as biotic or abiotic, and completing incomplete definitions of technical terms including respiration and decomposition. The post-test assessments also included questions that tested understanding of in-game concepts, presenting students with case studies that reported findings, requiring them to interpret the reported values from charts and tables and suggest the most likely explanation for the observations.

3.3 Feature Extraction

The game trace logs of the students comprise of 21 distinct action types, including *SpokeTo*, *Activated*, *MovedTo*, *ReceivedChatMessage* and *SentChatMessage*. At any given timestamp, we construct a 21-dimensional, count-based representation (i.e., the number of times a student performed each action type). These representations are

designed to support sequential predictions from the beginning of gameplay to the end with a goal to enable our stealth assessment models to make early predictions and inform adaptive interventions as early as possible. The counts for each action type were normalized using z-score normalization for effective modeling. In addition, pre-trained DistilBERT embeddings [22] were extracted for each message in the chat. DistilBERT, which was trained on Toronto Book Corpus and English Wikipedia (same as BERT), is a lighter version of BERT that maintains 95% of BERT's predictive performance while being 60% faster than BERT and 120% faster than ELMo and BiLSTM, lending itself to be a suitable technique for processing potential high volumes of chat messages exchanged by students in cases where our stealth assessment model is deployed to perform real-time assessments. The rubric for labeling the out-of-domain dataset is case-agnostic, so we use the uncased version of the DistilBERT embeddings to represent each chat message. Given the limited number of students in the dataset compared to the high dimensionality (768 dimensions) of the pre-trained DistilBERT embeddings, we perform principal component analysis (PCA) to reduce the dimensionality of the DistilBERT embeddings to 32 dimensions to be appropriate for training our stealth assessment models while preserving variance in the dataset.

3.4 Class Labeling

In our work, we cast each predictive task as a binary classification task so that our stealth assessment model can provide actionable feedback in the early phase of students' collaborative learning by achieving high predictive capacity. We define out-of-domain contribution of a student as the ratio of the number of out-of-domain messages sent by a student to the total number of out-of-domain messages sent in the group. If no out-of-domain messages were sent in the group, the out-of-domain contribution of all students in the group is considered to be zero. A higher score for the out-of-domain contribution metric indicates that the student negatively contributed to the group discussion by sending more out-of-domain chats. We performed median splits to determine high and low classification labels for each of the prediction tasks. The threshold for out-of-domain chat contribution was determined to be 20.3% (34 students in low category, 35 students in high category). Similarly, we performed a median split on the normalized post-test scores to determine binary class labels. The threshold used for labeling to post-test performance was 47.62% (33 students in low category, 36 students in high category).

4 Model Architecture

In this section, we present our multi-task learning stealth assessment model for predicting out-of-domain contribution to chat and post-test performance for students in a collaborative game-based learning environment. Based on a preliminary analysis of the data, our model is designed to make a prediction utilizing the past 20 actions performed by the student in the game, and 25 recent messages exchanged in the group chat for every new action that the student takes. The preliminary analysis suggested that using 20 actions and 25 recent messages as input to the stealth assessment models

not only captures useful game and dialogue context but also addresses data sparsity issues compared to utilizing the entire history of the actions and chats as input.

The model architecture embeds two recurrent neural networks as subnetworks to separately model game-trace logs and group chats, respectively. The subnetwork that processes game action features comprises a long short-term memory (LSTM) network (8 hidden units, 0.1 dropout, 0.01 L2 kernel, recurrent, and bias regularization factors), followed by a dropout layer with a dropout rate of 0.2. The subnetwork that processes recent group chats comprises a gated recurrent unit (GRU) network (8 hidden units, 0.1 dropout, 0.01 L2 kernel, recurrent, and bias regularization factors), followed by a dropout layer with a dropout rate of 0.2. The outputs from each of these subnetworks are concatenated to construct a combined representation (i.e., late fusion), which is fed into two separate dense layers (each with 1 output unit and sigmoid activation function) for predicting post-test performance and out-of-domain contribution, respectively. Our stealth assessment model was trained using the Adam optimizer [23] (learning rate = 0.01), and binary cross-entropy was used as the loss function. This model architecture is shown in Fig. 3. For single-task models, we only preserved layers specific to the prediction task, while removing all other layers added to the model architecture for the other prediction task. We maintain a non-neural random forest baseline that utilizes all available features to evaluate with our deep, multi-task learning-based stealth assessment model.

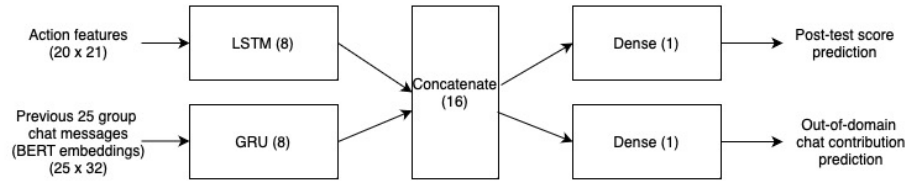

Fig. 3. Multi-task deep learning-based stealth assessment model architecture

5 Results

The model was evaluated using group-level 5-fold nested cross-validation. For each fold of cross-validation, we performed an 80-20 split on the dataset to create a training set and a separate validation set that is used to optimize the models. Accuracy and two early prediction metrics—standardized convergence point (SCP) [24] and convergence rate (CR) [25]—were used as evaluation metrics for our model. SCP measures how much in advance of the end of students' gameplay the model can consistently make accurate predictions about the outcome. The metric considers the average of the percentage of gameplay that elapses before a model can consistently and accurately predict the correct outcome, while penalizing non-converged sequences (i.e., student interaction sequences that did not yield a final correct prediction). CR indicates the percentage of converged student interaction sequences, which supplements SCP to demonstrate early prediction robustness of predictive models. In short, a higher CR and a lower SCP indicate a better early prediction performance.

The average of the results from each cross-validation fold is reported in this paper. We perform ablation studies on our stealth assessment model to determine the importance of game trace and group chat features for each of the prediction tasks and also determine if the prediction tasks benefit from multi-task learning of these two modalities. The results of our evaluation are presented in Tables 2 and 3 for post-test performance and out-of-domain contribution predictions, respectively.

Table 2. Single-task and multi-task learning results for predicting post-test performance. RF is the random forest baseline, DL is our deep learning-based stealth assessment model, A represents action features and G represents group chat features. (The best score is marked in bold for each metric.)

ML (Features)	Single-Task Learning				Multi-Task Learning			
	Macro Accuracy (%)	Micro Accuracy (%)	CR (%)	SCP (%)	Macro Accuracy (%)	Micro Accuracy (%)	CR (%)	SCP (%)
RF (A,G)	**67.5**	**66.5**	68.19	**52.26**	N/A	N/A	N/A	N/A
DL (A)	63.84	65.47	**70.84**	54.34	68.3	68.46	**70.68**	**49.53**
DL (G)	55.47	54.79	52.08	91.87	56.14	56.33	60.2	90.27
DL (A,G)	66.1	66.35	65	55.41	**70.37**	**71.01**	70.5	54.8

Table 3. Single-task and multi-task learning results for predicting out-of-domain performance. (The best score is marked in bold for each metric.)

ML (Features)	Single-Task Learning				Multi-Task Learning			
	Macro Accuracy (%)	Micro Accuracy (%)	CR (%)	SCP (%)	Macro Accuracy (%)	Micro Accuracy (%)	CR (%)	SCP (%)
RF (A,G)	63.34	62.78	69.86	59.08	N/A	N/A	N/A	N/A
DL (A)	**70.64**	**68.77**	71.53	**53.64**	67.67	68.51	74.21	49.44
DL (G)	46.39	46.38	41.54	96.13	47.34	47.42	41.34	88.14
DL (A,G)	68.03	67.36	**75.97**	54.22	**75.27**	**74.84**	**75.97**	**47.19**

6 Discussion

From Tables 2 and 3, we observe that our multi-task, deep learning-based stealth assessment model outperforms the random forest baseline for both prediction tasks, while the majority class baselines are 57.93% and 58.17% (macro accuracy) for post-test performance and out-of-domain contribution predictions, respectively. The best predictive accuracy is obtained using multi-task learning for predicting post-test performance (70.37% macro accuracy, 71.01% micro accuracy) and out-of-domain chat

contribution (75.27% macro accuracy, 74.84% micro accuracy) simultaneously, outperforming stealth assessment models based on single-task learning as well as random forest. The best early prediction results for both the post-test performance prediction task (49.53% SCP) and out-of-domain contribution prediction task (47.19% SCP) is also obtained in a multi-task learning setting. It should be noted that multi-task learning of game trace logs is especially effective for improving predictive performance of post-test performance, both in terms of accuracy and early prediction. This suggests that the out-of-domain contribution prediction task helped induce effective shared representations from the game trace logs subnetwork that is also beneficial for the post-test performance prediction task. From the early prediction results, we observe that the SCP for the out-of-domain contribution prediction task improves when there is multi-task learning with the multimodal data (SCP for the out-of-domain contribution prediction task in a single-task learning setting is 54.22%, compared to 47.19% when multi-task learning is applied). This could indicate that multi-task learning helps the model achieve stability in its predictions for out-of-domain contribution early, resulting in more consistent predictions with better convergence scores.

For single-task learning models for predicting post-test performance, we obtain the best result (66.1% macro accuracy, 66.35% micro accuracy, 65% convergence rate and 55.41% standardized convergence point) using a combination of game trace logs and group chat features, indicating that including group chat evidence is beneficial for the task of predicting post-test performance. Predictive accuracy of the single-task learning model for out-of-domain contribution is reduced when group chat features are used in addition to game trace log features, which suggests that the model is unable to effectively model recent group chat features when they are combined with game interaction features, whereas our multi-task stealth assessment models achieve the highest predictive performance as well as the best early prediction performance for out-of-domain contribution prediction, when the multimodal features are used.

7 Conclusion

Collaborative game-based learning environments are promising platforms for students to participate in team learning and collaborate on problem solving. Chat systems integrated in these environments serve as discussion forums and enable communication between groups of students. However, students may engage in out-of-domain discussions in the forum, which may be disruptive to the overall learning in the group. In this work, we present a deep learning-based stealth assessment model that supports early prediction of students' post-test performance and out-of-domain contribution in the group chat using multi-task learning. The multi-task learning-based stealth assessment models outperform the non-neural random forest baseline for both prediction tasks. We obtained the best prediction and early prediction results using both game trace logs and recent group chats as input features in a multi-task learning setting, suggesting that predicting post-test performance and out-of-domain contribution of students in collaborative game-based learning environments are related stealth assessment tasks that can benefit from multi-task learning.

In future work, it will be interesting to enrich chat message representations with sentiment scores, number of unique participants in the group chat context window, similarity of chat messages to in-game text, and the frequency of a student's contribution in recent chat history. Misspelled words in the chat can be corrected using a text normalization system, which could further improve robustness of stealth assessment models. Other pre-trained embeddings obtained from large language models can also be explored. Furthermore, it will also be important to investigate other modalities, such as posture and eye gaze data, for predicting learning outcomes and out-of-domain contribution of students in collaborative game-based learning environments. Finally, it will be important to incorporate the stealth assessment model alongside student-adaptive scaffolding to evaluate the impact on learning outcomes and processes within collaborative game-based learning environments.

Acknowledgements. This research was supported by the National Science Foundation under Grants DRL-1561655, DRL-1561486, IIS-1839966, and SES-1840120. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Dillenbourg, P., Järvelä, S., Fischer, F. The evolution of research on computer-supported collaborative learning. In Technology-enhanced learning (pp. 3-19). Springer, Dordrecht (2009).
2. Hmelo-Silver, C. E., Chernobilsky, E. Understanding collaborative activity systems: The relation of tools and discourse in mediating learning. In Embracing Diversity in the Learning Sciences: Proceedings of the Sixth International Conference of the Learning Sciences (p. 254). Psychology Press (2004, October).
3. Jeong, H., Hmelo-Silver, C. E., Jo, K. Ten years of computer-supported collaborative learning: A meta-analysis of CSCL in STEM education during 2005–2014. Educational research review, 28, 100284 (2019).
4. Engle, R. A., & Conant, F. R. Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of learners classroom. Cognition and instruction, 20(4), 399-483 (2002).
5. Saleh, A., Chen, Y., Hmelo-Silver, C. E., Glazewski, K. D., Mott, B. W., & Lester, J. C. Coordinating scaffolds for collaborative inquiry in a game-based learning environment. Journal of research in science teaching, 57(9), 1490-1518 (2020).
6. Park, K., Sohn, H., Mott, B., Min, W., Saleh, A., Glazewski, K., Hmelo-Silver, C., Lester, J. Detecting disruptive talk in student chat-based discussion within collaborative game-based learning environments. In LAK21: 11th International Learning Analytics and Knowledge Conference (pp. 405-415) (2021, April).
7. Jeong, H., Hmelo-Silver, C. Technology supports in CSCL (2012).
8. Carpenter, D., Emerson, A., Mott, B. W., Saleh, A., Glazewski, K. D., Hmelo-Silver, C. E., & Lester, J. C. Detecting off-task behavior from student dialogue in game-based collaborative learning. In International Conference on Artificial Intelligence in Education (pp. 55-66). Springer, Cham (2020, July).

9. Sabourin, J. L., Rowe, J. P., Mott, B. W., & Lester, J. C. Considering Alternate Futures to Classify Off-Task Behavior as Emotion Self-Regulation: A Supervised Learning Approach. Journal of Educational Data Mining, 5(1), 9-38 (2013).
10. Mislevy, R. J., Steinberg, L. S., Almond, R. G. Focus article: On the structure of educational assessments. Measurement: Interdisciplinary research and perspectives, 1(1), 3-62 (2003).
11. Shute, V. J. Stealth assessment in computer-based games to support learning. Computer games and instruction, 55(2), 503-524 (2011).
12. Henderson, N., Acosta, H., Min, W., Mott, B., Lord, T., Reichsman, F., Dorsey, C., Wiebe, E., Lester, J. Enhancing Stealth Assessment in Game-Based Learning Environments with Generative Zero-Shot Learning. International Educational Data Mining Society (2022).
13. Kim, Y. J., Almond, R. G., Shute, V. J. Applying evidence-centered design for the development of game-based assessments in physics playground. International Journal of Testing, 16(2), 142-163 (2016).
14. Zhao, W., Shute, V., Wang, L. Stealth assessment of problem-solving skills from gameplay. In Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC),(15212) (2015).
15. Shute, V. J., Rahimi, S. Stealth assessment of creativity in a physics video game. Computers in Human Behavior, 116, 106647 (2021).
16. Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Smith, A., Wiebe, E., Boyer, K. E., Lester, J. C. DeepStealth: Game-based learning stealth assessment with deep neural networks. IEEE Transactions on Learning Technologies, 13(2), 312-325 (2019).
17. Zhang, Y., Yang, Q. A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering (2021).
18. Gupta, A., Carpenter, D., Min, W., Rowe, J. P., Azevedo, R., Lester, J. C. Multimodal Multi-Task Stealth Assessment for Reflection-Enriched Game-Based Learning. In MAIED@ AIED (pp. 93-102) (2021).
19. Dillenbourg, P., Fischer, F. Computer-supported collaborative learning: The basics. Zeitschrift für Berufs-und Wirtschaftspädagogik, 21, 111-130 (2007).
20. Sung, H. Y., Hwang, G. J. A collaborative game-based learning approach to improving students' learning performance in science courses. Computers & education, 63, 43-51 (2013).
21. Carpenter, D., Emerson, A., Mott, B. W., Saleh, A., Glazewski, K. D., Hmelo-Silver, C. E., Lester, J. C. Detecting off-task behavior from student dialogue in game-based collaborative learning. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I 21* (pp. 55-66). Springer International Publishing (2020).
22. Sanh, V., Debut, L., Chaumond, J., Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019).
23. Kingma, D. P., Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
24. Min, W., Baikadi, A., Mott, B., Rowe, J., Liu, B., Ha, E. Y., Lester, J. A generalized multidimensional evaluation framework for player goal recognition. In Twelfth Artificial Intelligence and Interactive Digital Entertainment Conference (2016, September).
25. Blaylock, N., Allen, J. Corpus-based, statistical goal recognition. In IJCAI (Vol. 3, pp. 1303-1308) (2003, August).