CPSCoach: The Design and Implementation of Intelligent Collaborative Problem Solving Feedback

Angela E.B. Stewart[1][0000-0002-6004-9266], Arjun Rao[2][0009-0007-0734-3822], Amanda Michaels[3][0009-0006-7644-5331] , Chen Sun[4][0000-0002-7575-5091] , Nicholas D. Duran[3][0000-0002-8872-5617] , Valerie J. Shute[5][0000-0002-9179-017X] , and Sidney K. D'Mello[2][0000-0003-0347-2807]

[1] University of Pittsburgh, Pittsburgh PA 15260, USA
[2] University of Colorado Boulder, Boulder CO 80309, USA
[3] Arizona State University, Glendale AZ 85306, USA
[4] University of Manchester, Manchester, United Kingdom
[5] Florida State University, Tallahassee FL 32306, USA
angelas@pitt.edu

Abstract. We present the design of CPSCoach, a fully-automated system that assesses and provides feedback on collaborative problem solving (CPS) competencies during remote collaborations. We leveraged existing data to develop deep NLP models that automatically assess the CPS competencies from speech, achieving moderate to high accuracies (average area under the receiver operating characteristic curve of .78). We engaged 43 participants in an iterative process to design the feedback mechanism, resulting in the first prototype of CPSCoach. We conducted a user study with 20 dyads who engaged with CPSCoach over multiple rounds. Participants thought the system was usable, but they were mixed about the accuracy of the feedback. We discuss design considerations for feedback systems aimed at improving CPS competencies.

Keywords: collaborative problem solving, intelligent feedback.

1 Introduction and Related Work

Collaborative problem solving (CPS) is a ubiquitous phenomenon, occurring when multiple people engage in a coordinated attempt to solve a problem [1]. Productive CPS involves combining socio-cognitive processes, such as maintaining a shared vision of the problem and coordinating teammate strengths to implement a solution [2]. Remote CPS is increasingly common [3]. However, remote collaborations are especially difficult, as the social signals from face-to-face communication are muted or non-existent [3]. Currently, students are expected to master CPS competencies through project-based group learning [2], yet they are evaluated on project outcomes, not CPS skill [2]. Further, they do not receive meaningful feedback on their CPS skills [2], making it difficult for them to acquire and improve these competencies.

Software systems do exist that support collaborative skills, although these largely are task-specific (e.g., interfaces to enhance team communication programming [4]). Most similar to our work are systems that support awareness of behavior in collaborative interactions. In this literature, low-level behaviors are synthesized into interpretable metrics, and visualized. For example, real-time visualizations of turn-taking behaviors (e.g., interruptions) have been aggregated from speech signals [5].

Taken together, there is a conspicuous absence of pathways to CPS proficiency. Accordingly, we developed CPSCoach, a system that uses NLP to automatically assess CPS skills [1] and provide feedback between short collaborations in an educational physics game. CPSCoach is based in theories on the importance of personalized and immediate feedback to support learning [6]. We extend current collaboration feedback by focusing on high-level CPS constructs, rather than low-level behaviors. Low-level behaviors are related to CPS, but CPS is far more than this. For example, ample verbal participation is related to problem solving ideation. However, this is insufficient as the team could be entirely off topic or discussing unproductive ideas. Accordingly, we move from low-level concrete and generic behaviors (e.g., amount of verbal participation) to abstract and specialized CPS behaviors (e.g., joint knowledge construction). In this work, our goal is to identify key lessons from a proof-of-concept prototype to inform future versions of CPSCoach.

2 Intervention Design and User Study

We conducted an iterative design study to explore effective designs for CPS feedback, focusing on instructional content, translation of model predictions to interpretable metrics, and the user interface. Participants in the design study were 43 college students from a large public university (female = 69%; male = 29%; 2% = non-binary; Asian = 23%, Black = 3%; Hispanic = 9%, white = 60%).

Our final prototype used a theoretically-grounded, empirically-validated CPS framework [1] which consists of three facets: shared knowledge construction, negotiation/coordination, and maintaining team function. This framework was applied in the context of Physics Playground [7], a physics problem-solving game. After collaborating in Physics Playground, teammates viewed their feedback in CPSCoach. The scores overview page included their scores for the current round and past scores. Participants could view a facet-specific page, which included a definition and indicative behaviors, and annotated videos of teams engaging in the behaviors.

Our CPS facet machine learning models were trained on a dataset of 94 triads playing Physics Playground [8]. We used the Bidirectional Encoder Representations from Transformers (BERT) model [9], with pre-trained word embeddings that were fine-tuned to our data. We chose hyper-parameters based on recommendations from [9] (e.g., fine-tuning over four epochs, batch size = 32, sequences padded or truncated at 300 words). We used team-level 10-fold cross validation such that all utterances from a team were in the training set, or testing set, but never both. Using this approach, we achieved area under the receiver operating characteristic curve (AUROC) values of .88, .83, and .82 on our test data for shared knowledge construction, negotiation/coordination, and maintaining team function, respectively (chance = 0.5). To use the model for real-time assessment, we retrained the model with all the data, preserving the hyper-parameters. The final model yielded one probability score per facet that a given utterance was a positive example of that facet.

The probability scores were converted into interpretable metrics. Our scoring scheme used a normed-reference approach based on the training data. Given there are multiple utterances per CPS round, we averaged prediction probabilities across the round (per individual, per facet). We fit a probability density function on that average, after Winsorizing [10] the top and bottom 1.25% of outliers. We only used utterances in the training dataset from the first 10 minutes of the interaction as only 10 minutes of data is used in our CPSCoach. A final numeric score was determined by the cumulative distribution function where a score that is high, compared to the reference group, yields a higher value (range 0% to 100%). We also displayed ordinal feedback (low – bottom 25%, average – middle 50%, high – top 25%).

We recorded audio using Zoom, which records separate audio tracks for each teammate. After recording, we transcribed each participant's audio using IBM Watson [11] and the resultant utterances were submitted to the BERT model. We used only the first ten (of 15) minutes of a CPS interaction due to processing latency.

We conducted a user study with 40 students (20 dyads) from two large public universities in the United States (School One – N = 32, female = 55%, male = 42%, non-binary = 3%, Asian = 29%, Hispanic = 3%, white = 61%, other race = 3%; School Two – N = 8, female = 75%, male = 25%, Asian = 25%, Hispanic = 13%, Native American or Pacific Islander = 13%, white = 38%, other race = 13%). Participants completed a game tutorial and warmup, followed by a video orienting them to the upcoming task, CPS framework, and feedback. Teammates engaged in three to four rounds of collaborative gameplay and feedback, as time permitted (10 – 15 minutes each). They were given up to seven minutes to individually review feedback between rounds. They then completed a short survey on their perception of the accuracy of the scores, and usefulness of the tips for improving the scores. For example, they were asked: "Your score for sharing ideas and expertise was [score]%. How accurate was this score?" At the end of the study, participants completed the System Usability Scale [12], as a measure of general usability ($\alpha = .66$). All measures were completed on a five-point Likert scale. To conclude the study, we conducted semi-structured interviews to further understand participant's perception of CPSCoach. Other measures not germane to the present study were also collected.

3 Results and Discussion

To understand behavioral patterns as participants interacted with the feedback system, we examined average time spent on each element of CoachCPS (scores overview, facet text, and facet videos). We conducted two-tailed paired-samples t-tests between each round. Participants spent a similar amount of time on the score overview page across all three rounds (average of 52.84s, $p > .05$). However, for the facet text, participants spent a third of the time on each subsequent round (round 1 = 61.86s, round 2 = 41.44s, round 3 = 21.99s; $p < .05$ for round 1 to 2, $p < .01$ for round 2 to 3, $p < .001$ for round 1 to 3). There was a similar finding for the facet videos (round 1 = 159.29s, round 2 = 110.55s, round 3 = 42.88s) where there were significant differences between rounds 1 and 3 ($p < .001$), and 2 and 3 ($p < .01$).

We also examined participants' subjective perceptions of the feedback system. First, we considered ratings of score accuracy and usefulness. For each participant, we averaged across all rounds, to obtain an overall accuracy and usefulness measure. We computed an overall usability rating for the CPS system as the average of items from the System Usability Scale. Descriptive statistics are shown in **Table 1**. We also conducted a thematic analysis [13] of the interviews, in order to gain a deeper understanding of participants' perceptions of CPSCoach.

Participants found the models to be somewhat accurate (a mean of 3.4 across all facets). A theme in the participant interviews was doubts about model accuracy, with 10% mentioning the accuracy of the shared knowledge construction and maintaining team function facets, and 10% expressing general doubt about score accuracy. For example, one user noted "I was kind of surprised that the sharing ideas and expertise score was pretty low for the first two rounds. I thought that we were sharing ideas pretty well for those first two rounds and it kind of surprised me that they were in the low area, especially when low is the bottom quarter." Participants also found the facet text and videos to be somewhat useful (average of 3.64 across facets). This sentiment was also reflected in the interviews, with 18% of participants mentioning they found the video examples helpful. For instance, one user noted: "Hearing the examples of what other people did I feel like was more helpful not so much reflecting on what I did last time, but what I can do next time instead." Ratings of usability of the feedback system were high (average of 4.24 out of 5), which was reflected in the interviews, as 20% of participants noted that the feedback was clear and easy-to-use.

Participants had suggestions for improving CoachCPS, such as examples of good and bad CPS from their own behavior (13%). 20% indicated they wanted the facet tips and videos to change each round, which gives them more examples to learn from, and 10% wanted indicator-level feedback. As an illustrative example of these two points, one participant made suggestions for how to provide indicator level feedback that changed each round: "I know it gives you the percentages, but in like positive team building, it said you did a really good job doing X but you could work on Y, something like that. So it feels like it's more based on the round instead of just the same ideas to improve every time…like within positive work environment asking for collaboration ideas you could say like oh like you did a good job asking for ideas but you could have done better with like specific solutions and stuff like that."

Table 1. Descriptive statistics for subjective perception measures.

Measure	M	SD	Min	Max
Construction of Shared Knowledge Accuracy	3.40	.83	1	5
Construction of Shared Knowledge Usefulness	3.58	1.0	1	5
Negotiation/Coordination Accuracy	3.70	.74	1.67	5
Negotiation/Coordination Usefulness	3.76	1.02	1	5
Maintaining Team Function Accuracy	3.05	.84	1.33	4.33
Maintaining Team Function Usefulness	3.58	1.03	1	5
System Usability Scale	4.24	.62	2.4	5

Our user study led us to two key lessons learned. First, *feedback should change each time a user views it, to encourage them to continuously improve their CPS.* We found that participants spent less time each round interacting with CPSCoach. This finding could be linked to increased familiarity with the content as rounds progressed. However, in interviews, participants requested different video examples each round, which could increase engagement and in turn support score improvement. Further feedback personalization, such as video examples extracted from the previous interaction, might aid engagement. This kind of personalization poses both technical and design challenges. Specific examples from the previous round of CPS require accurate models at fine-grained time intervals (e.g., 30s rather than aggregated over a ten minutes). It is an open question as to how to automatically select examples from the interaction that are meaningful and can prompt positive behavior change

Our second lesson was that *increased transparency of the CPS feedback and underlying scoring mechanisms could increase users' trust in the feedback accuracy.* Participants rated usability of CPSCoach highly, but ratings of score accuracy were average. Facets with the two lowest accuracy ratings (shared knowledge construction and maintaining team function) were specifically mentioned in the interviews. Thus, there is a need for increased trust and transparency in the feedback. In order for participants to buy in to an AI system, the outcome of the system must be reasonably predictable [14]. Perhaps participants being surprised by the score resulted in the feedback being less effective. Additional training on the nuances of the CPS framework could help boost participant trust, by differentiating the facets they are being scored on from preconceived notions. Feedback designs proposed by participants (e.g., examples from their interactions) could also increase transparency of the underlying CPS model and in turn trust in the feedback.

Given that this is a first prototype CPSCoach, there are limitations that should be addressed. The purpose of this work was to design a prototype feedback system and conduct an initial user study. Accordingly, our study design did not include a control group, or formal efficacy evaluation. Findings from this study indicated several areas of improvement, thus it is prudent to make these changes before conducting an efficacy study. As collaborations increasingly move online, now, more than ever, we need to support teams CPS. Our work presents important steps towards this goal of developing systems that equip teammates with the skills they need to become more effective collaborative problem solvers.

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805. The opinions expressed are those of the authors and do not represent views of the NSF.

References

1. Sun, C., Shute, V.J., Stewart, A.E.B., Yonehiro, J., Duran, N., D'Mello, S.K. Towards a generalized competency model of collaborative problem solving. Comput Educ. 143, 103672 (2020). https://doi.org/https://doi.org/10.1016/j.compedu.2019.103672.

2. Graesser, A.C., Fiore, S.M., Greiff, S., Andrews-Todd, J., Foltz, P.W., Hesse, F.W.: Advancing the Science of Collaborative Problem Solving. Psychological Science in the Public Interest. 19, 59–92 (2018). https://doi.org/10.1177/1529100618808244.
3. Schulze, J., Krumm, S.: The “virtual team player”: A review and initial model of knowledge, skills, abilities, and other characteristics for virtual collaboration. Organizational Psychology Review. 7, 66–95 (2017). https://doi.org/10.1177/2041386616675522.
4. Čubraniundefined, D., Storey, M.A.D., Čubranić, D., Storey, M.A.D.: Collaboration Support for Novice Team Programming. In: Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work. pp. 136–139. Association for Computing Machinery, New York, NY, USA (2005). https://doi.org/10.1145/1099203.1099229.
5. Faucett, H.A., Lee, M.L., Carter, S.: I Should Listen More: Real-time Sensing and Feedback of Non-Verbal Communication in Video Telehealth. Proc. ACM Hum.-Comput. Interact. 1, 44:1--44:19 (2017). https://doi.org/10.1145/3134679.
6. Shute, V.J.: Focus on Formative Feedback. Rev Educ Res. 78, 153–189 (2008). https://doi.org/10.3102/0034654307313795.
7. Shute, V.J., Smith, G., Kuba, R., Dai, C.-P., Rahimi, S., Liu, Z., Almond, R.: The Design, Development, and Testing of Learning Supports for the Physics Playground Game. Int J Artif Intell Educ. 31, 357–379 (2021). https://doi.org/10.1007/s40593-020-00196-1.
8. Stewart, A.E.B., Keirn, Z., D’Mello, S.K.: Multimodal modeling of collaborative problem-solving facets in triads. User Model User-adapt Interact. 31, 713–751 (2021). https://doi.org/10.1007/s11257-021-09290-y.
9. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: HuggingFace’s Transformers: State-of-the-art Natural Language Processing, (2019).
10. Dixon, W.J., Yuen, K.K.: Trimming and winsorization: A review. Statistische Hefte. 15, 157–170 (1974). https://doi.org/10.1007/BF02922904.
11. IBM, https://www.ibm.com/watson/services/speech-to-text/, last accessed 2018/05/01.
12. Brooke, J., others: SUS-A quick and dirty usability scale. Usability evaluation in industry. 189, 4–7 (1996).
13. Blandford, A., Furniss, D., Makri, S.: Qualitative HCI Research: Goaing Behind the Scenes. Morgan & Claypool (2016). https://doi.org/10.2200/S00706ED1V01Y201602HCI034.
14. Bostrom, N., Yudkowsky, E.: The ethics of artificial intelligence. The Cambridge handbook of artificial intelligence. 1, 316–334 (2014).