Chapter 24 – Assessment of Collaborative Problem Solving

Arthur C. Graesser[1], Zhiqiang Cai[1], Xiangen Hu[1], Peter W. Foltz[2], Samuel Greiff[3], Bor-Chen Kuo[4], Chen- Huei Liao[4], and David Williamson Shaffer[5]
University of Memphis[1], University of Colorado Boulder and Pearson[2], University of Luxembourg[3], National Taichung University of Education[4], University of Wisconsin[5]

Introduction

Collaborative problem solving (CPS) is one of the important 21st century skills that has attracted interest in international assessments, national assessments of middle and high school students, colleges, business, and the military (Griffin & Care, 2015; Hesse, Care, Buder, Sassenberg & Griffin, 2015; NRC, 2011; OECD, 2013; Sottilare et al., 2015). CPS is an essential skill in the home, the workforce, and the community because much of the planning, problem solving, and decision making in the modern world is performed by teams. The success of a team can be threatened by a social loafer, an uncooperative unskilled member, or a counterproductive alliance, whereas it can be facilitated by a strong team member that draws out different perspectives, helps negotiate conflicts, assigns roles, promotes team communication, and guides the team to overcome troublesome obstacles (Fiore, Wiltshire, Oglesby, O'Keefe & Salas, 2014; Salas, Cooke & Rosen, 2008).

CPS differs from individual problem solving (IPS) in ways that may have both positive and negative consequences. CPS allegedly has advantages over IPS because 1) there is a more effective division of labor, 2) the solutions incorporate information from multiple sources of knowledge, perspectives, and experiences, and 3) the quality of solutions is stimulated by ideas of other team members. There are also potential disadvantages of CPS to the extent that 1) team members waste time with irrelevant discussion, 2) there is diffusion of responsibility in completing tasks, and 3) disagreements among team members occur that paralyze progress in solving the problem.

At the international level, CPS was selected by the Organisation for Economic Co-operation and Development (OECD) as a new development for the Program for International Student Assessment (PISA) in the 2015 international survey of student skills and knowledge (Graesser, Forsyth & Foltz, 2016; OECD, 2013, 2015). Fifteen-year-old students from over three dozen countries completed this PISA CPS 2015 assessment in addition to assessments of mathematics, science, literacy, and other proficiencies. One of the goals of this chapter is to describe how CPS was assessed in PISA CPS 2015.

PISA used computer agents in the 2015 assessment. That is, a single human interacts with one, two, or three computer agents as team members rather than other humans. Conversation-based assessments with computer agents are manifested by chat conversations as well as actions of team members (Zapata-Rivera, Jackson & Katz, 2015). Computer agents are believed to provide control over the social interaction so that important assessments can be made with consistency and control, two requirements that communicating with fellow humans could not provide. Agents also provide control over logistical and measurement problems that stem from 1) assembling groups of humans (via computer mediated conversation) in a timely manner, 2) the necessity of having multiple teams per student to obtain reliable assessments in different circumstances, and 3) extreme measurement error when particular students are paired with other humans who do not collaborate well. A second goal of this chapter is to describe how agents can be used to provide meaningful assessments of CPS.

Although conversation-based assessments with agents can provide meaningful assessments of CPS, there is still an important goal of assessing interactions among humans. That requires an automated analysis of natural language and discourse in addition to identifying how particular problem-solving patterns map onto important CPS proficiencies (e.g., establishing shared knowledge, taking initiative, communicating important information to the group). The third goal of this chapter is to identify some of the automated approaches that show promise in automated assessments of CPS among humans. These methods could be integrated with the Generalized Intelligent Framework for Tutoring (GIFT) in future developments by piggybacking on and expanding existing applications of natural language processing in GIFT.

Related Research

There have been a number of theoretical frameworks for analyzing CPS. Some of the prominent ones include the Center for Research on Evaluation, Standards, and Student Testing's (CRESST) teamwork processing model (O'Neil, Chuang & Baker, 2010), the teamwork models of Salas, Fiore, and colleagues (Fiore et al., 2010; Salas, Cooke & Rosen, 2008) and the Assessment and Teaching of 21st Century Skills (ATC21S; Griffin & Care, 2015; Hesse et al., 2015). All of these frameworks have both a *cognitive* dimension that includes problem solving and other cognitive processes and a *collaborative* dimension that includes communication and other social interaction processes. These approaches were incorporated in PISA CPS 2015 (Graesser et al., 2016; OECD), the framework under direct focus in this chapter.

The problem-solving dimension in PISA CPS 2015 framework incorporated the same PISA 2012 problem solving framework that targeted individual problem solving (Funke, 2010; OECD, 2010; Greiff, Kretzschmar, Müller, Spinath & Martin, 2014). There were four cognitive processes (or competencies) on the problem-solving dimension:

1) **Exploring and understanding.** Interpreting the initial information about the problem and any information that is uncovered during the course of exploring and interacting with the problem.
2) **Representing and formulating**. Identifying global approaches to solving the problem, relevant strategies and procedures, and relevant artefacts (e.g., graphs, tables, formulae, symbolic representations) to assist in solving the problem.
3) **Planning and executing.** Constructing and enacting goal structures, plans, steps, and actions to solve the problem. The actions can be physical, social, or verbal.
4) **Monitoring and reflecting**. Tracking the steps in the plan to reach the goal states, marking progress, and reflecting on the quality of the progress or solutions.

There were three processes on the collaborative dimension:

1) **Establishing and maintaining shared understanding**. Keeping track of what each other knows about the problem (i.e., shared knowledge, common ground; Clark, 1996), the perspectives of team members, and a shared vision of the problem states and activities (Cannon-Bowers & Salas, 2001; Dillenbourg & Traum, 2006).
2) **Taking appropriate actions to solve the problem.** Performing actions that follow the appropriate steps to achieve a solution. This includes physical actions and communication acts that advance the solution to the problem.
3) **Establishing and maintaining group organization.** Helping organize the group to solve the problem by considering the talents and resources of group members during the assignment of roles; following the rules of engagement for one's own roles as well as handing obstacles to tasks assigned to other team members.

When the 4 problem-solving processes are crossed with the 3 collaboration processes, there are 12 skills in the resulting *CPS assessment matrix*. Table 1 shows this matrix that was adopted in the PISA CPS 2015 framework. A satisfactory assessment of CPS would assess the skill levels of students for each of these 12 cells and these would contribute to a student's overall *CPS proficiency measure*.

Table 1. Copied from OECD (2013). PISA 2015 collaborative problem solving framework.

	(1) Establishing and maintaining shared understanding	**(2) Taking appropriate action to solve the problem**	**(3) Establishing and maintaining team organization**
(A) Exploring and Understanding	(A1) Discovering perspectives and abilities of team members	(A2) Discovering the type of collaborative interaction to solve the problem, along with goals	(A3) Understanding roles to solve problem
(B) Representing and Formulating	(B1) Building a shared representation and negotiating the meaning of the problem (common ground)	(B2) Identifying and describing tasks to be completed	(B3) Describe roles and team organization (communication protocol/rules of engagement)
(C) Planning and Executing	(C1) Communicating with team members about the actions to be/ being performed	(C2) Enacting plans	(C3) Following rules of engagement, (*e.g.,* prompting other team members to perform their tasks.)
(D) Monitoring and Reflecting	(D1) Monitoring and repairing the shared understanding	(D2) Monitoring results of actions and evaluating success in solving the problem	(D3) Monitoring, providing feedback and adapting the team organization and roles

As mentioned, the PISA CPS 2015 assessment had students interact with computer agents rather than other humans. The following definition of CPS was articulated in the PISA CPS 2015 framework (OECD, 2013: *Collaborative problem-solving competency is the capacity of an individual to effectively engage in a process whereby* ***two or more agents*** *attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution.* An agent could be either a human team member or a computer agent that interacts with the student. The final assessment that was adopted had students interact with one to three computer agents instead of other humans. Therefore, the overall CPS proficiency measure assessed how well a single human interacted with computer agents during the course of problem solving. The computer agents were minimalist agents in a chat facility, without text-to-speech, animation, or visual depictions of what they looked like. This was necessary to eliminate biases of culture, personality, and emotions, which were beyond the scope of the PISA CPS assessment.

A central advantage of assessments with computer agents is the degree of control over the conversation. The discourse contributions of the two agents (A1, A2) and any associated digital media (M) can be coordinated so that each [A1, A2, M] sequential display is functionally a single episodic unit (U) to which the human responds through language, action, or silence in a particular human turn (HT). Thus, there is an orchestrated finite-state transition network that alternates between episodic units (U) and human turns (HT), which is formally isomorphic to a dialogue. This is very different than a collaboration in which many people can speak simultaneously and overlap in time (Clark, 1996). Conditional branching can occur in the state-transition network (STN) so that the computer's generation of U_{n+1} at turn n+1 is contingent on the state of the human turn HT_n at turn n. However, the degree of branching was limited to a small number of states associated with each human turn (HT_n) in PISA CPS 2015; there were 2–4 alternative multiple-choice options at each turn (i.e., either chat options or alternative actions to be performed). Consequently, the fan out of conditional branching was not complex and the turn-taking frequently converged at points of assessment rather than diverging in many directions. Only one score was associated with each episodic unit and each episodic unit was aligned with one and only one of the 12 cells in the CPS assessment matrix.

The design of the PISA CPS 2015 assessment was compatible with the normal psychometric modeling in the world of assessment, where multiple-choice tests are ubiquitous. Traditional psychometric assessments routinely include a fixed set of items (i.e., episodic units) that all humans experience. Analogously, PISA CPS 2015 had a fixed sequence episodic units (U_1, U_2, … U_m) that occurred at specific points as the problem was solved and the responses of the human were automatically recorded (as clicks on action options or chat options). The conversations were designed so that the conversations would naturally close shortly after the human responded to an episodic unit and the subsequent episodic unit was launched (e.g., "Thanks for your input, let's go on"). Assessment scores were collected for each student for the M episodic units that collectively covered each of the 12 cells in the CPS assessment matrix. These scores contributed to overall CPS proficiency measures that have not yet been finalized by OECD.

Students encounter a diverse set of situations in PISA CPS 2015 in order to make sure that important conditions are covered in the assessment. Students who respond randomly to the response options would obviously receive low values on CPS proficiency as well as the collaboration and problem solving dimensions. A student may be a good team player and be responsive, but not take the initiative when there are problems (e.g., an agent who is unresponsive, or a new obstacle in the problem occurs). A student may take some initiative when there are breakdowns, but not be able to handle very complex cognitive problems. A student who scores high in CPS proficiency takes the initiative in moving the team to achieve group goals during difficult times (conflicts, incorrect actions, unresponsive team members) and can also handle complex problems with many cognitive components that burden working memory and require reasoning. Episodic units for all of these situations are needed in order to have an adequate CPS assessment. In contrast, many of these situations might not arise when a student interacts with other humans so there would be missing scores for some of the 12 cells.

Computer agents may be suitable for providing a summative assessment of CPS proficiency that is both reliable and valid. Available data have so far supported the validity of the PISA CPS 2015 framework. For example, a factor analysis has shown an extremely high correspondence between a human-agent CPS assessment and a human-human assessment in a sample in Germany students (Greiff, personal communication). Kuo et al. (2016) conducted an assessment in Taiwan that adopted the PISA CPS 2015 assessment framework. The study developed an internet-based CPS assessment with conversational agents on five tasks to be completed in 100 minutes. There were over 50,000 ninth and tenth grade students who participated between October 2014 and February 2015. The problem-solving dimension in the PISA CPS 2015 assessment showed a similar ordering of competencies for the four problem-solving components (A > B > C > D) as were found for the PISA 2012 assessments of individual problem solving. Although the complete data for PISA CPS 2015 is still being analyzed for over 400,000 students in three to four dozen different countries, the reliability of the data in field trials is encouraging.

Discussion

Although computer agents may be suitable for a summative assessment of CPS proficiency, there are major limitations with this approach for teams of humans and formative CPS assessment. Computer-based environments for teams (whether they be collaborative learning, problem solving, or work) need automatic tracking and analysis of the language, actions, and social interactions of human team members. Computer-based environments need to adaptively, intelligently, and immediately respond to the team members based on the automated assessments of CPS proficiencies and many other cognitive and noncognitive characteristics of team members. The agent-based assessment in PISA CPS 2015 does not offer any help in developing a computer environment for tracking and responding to teams of humans. The latter would be needed in GIFT (Sottilare et al, 2017).

The remainder of this chapter identifies some promising ways of automatically tracking the language and discourse of humans in team chat interactions. Ideally, we would be able to map particular language and conversation patterns onto the cells of CPS assessment matrix. If these patterns could be detected automatically, then there is a principled theoretical foundation for 1) a formative assessment of CPS skills of team members and 2) recommendations on how the computer environment should respond to unproductive teams or team members.

A community of researchers in the learning sciences and computational linguistics have investigated conversations in small groups by analyzing the log files of computer-mediated interactions in chat and discussion forums (Dowell, Graesser & Cai, in press; Foltz & Martin, 2008; Liu, Von Davier, Hao, Kyllonen & Zapata-Rivera, 2015; Mu, Stegmann, Mayfield, Rosé & Fischer, 2012; Shaffer, Collier & Ruis, in press; Tausczik & Pennebaker, 2013; Von Davier & Halpin, 2015). The conversations have been analyzed by a variety of automated text analysis tools, such as state-transition networks that track speech acts of team players (Morgan, Keshtkar, Duan & Graesser, 2012), latent semantic analysis (Foltz & Martin, 2008; Gorman et al., 2003), epistemic network analysis (Shaffer et al., 2009), Coh-Metrix (Graesser, McNamara, et al., 2014), and Linguistic Inquiry and Word Count (Pennebaker, Booth & Francis, 2007). These automated tools have been applied to conversations in their entirety, to subsets of the conversation at a particular window size (e.g., 5 turns in a row), to single conversational turns, to adjacent conversational turns, and to turns of specific team members. The conversation profile includes measures of team cohesion, percentage of on-topic versus off-topic contributions, amount of new information, characteristics of team members (e.g., driver, follower, social loafer), alliances between team members, and presence of specific conversation patterns. It is beyond the scope of this chapter to describe in detail these automated approaches (see Graesser, Dowell, Clewley & Shaffer, submitted), but we do highlight some examples to illustrate the prospects of this approach.

Matches to Expectations

In many applications of team problem solving, there are a set of *expectations* that need to be covered to solve the problem. An expectation is a sentence, clause, proposition, or expression of comparable length, as discussed in reports on AutoTutor (Graesser, 2016; Cai, Graesser & Hu, 2015). A solution to a problem consists of a set of expectations that hopefully would be covered by the team. The team or team member received higher scores to the extent that more expectations are articulated during the chat conversation. Physical actions are also handled in this way: performance increases as more critical actions are performed.

Advances in computational linguistics and semantics have made impressive gains in the accuracy of semantic matches between one short text (i.e., a sentence or two) and another short text (Rus, Lintean, Graesser & McNamara, 2012; Rus & Ştefănescu, 2016). The accuracy is not always perfect, but it often is impressive and on par with human experts who judge the semantic similarity of pairs of short texts. The AutoTutor research team has evaluated many semantic matchers over the years in AutoTutor and other intelligent tutoring systems (ITSs) with conversational agents (Cai et al., 2011; Graesser, Penumatsa, Ventura, Cai & Hu, 2007; Rus et al., 2012). The semantic matchers automatically compute the semantic similarity between a student's verbal contribution and an expectation, with a similarity score that varies from zero to one. These semantic match algorithms have included keyword overlap scores, word overlap scores that place higher weight on lower frequency words in the English language, scores that consider the order of words, latent semantic analysis cosine values, comparisons to regular expressions, and procedures that compute semantic logical entailment. As an example, Cai et al. (2011) reported that the correlation of similarity scores between AutoTutor and human expert judges was $r = 0.667$, about the same as between two trained judges ($r = 0.686$). Interestingly syntactic parsers did not prove useful in these analyses because a high percentage of the students' contributions are vague, telegraphic, elliptical, and ungrammatical. At the time of this writing, the best automated semantic matcher is the Semantic Similarity (SEMILAR) system developed by Rus et al. (2013). SEMILAR won the semantic textual similarity competition at *SemEval-2015*, the premier international forum for semantic evaluation.

Matches to expectations are powerful in assessments of CPS to the extent that the solutions to a problem are known ahead of time, as in the case of PISA CPS 2015. Indeed, there could be a set of expectations associated with each of the 12 cells in the CPS assessment matrix and these could be scored for each team member over the course of the CPS interaction. Unfortunately, this approach does not work when there are no expectations in a CPS application. The subsequent approaches can be applied when a problem does not have a finite set of associated expectations.

Automated Speech Act Classification and State Transition Networks

The content of each turn is classified into speech acts and each speech act is assigned to a category (Liu et al., 2015; Morgan et al., 2012). For example, the speech act categories defined by Rus, Graesser, Moldovan, and Niraula (2012) were Statement, Question, Request, Reaction, MetaStatement, Expressive Evaluation, and Greeting. Automated speech act classification has achieved a moderate degree of accuracy compared with trained human annotators (Olney et al., 2003; Rus et al., 2012; Samei et al., 2014). A chat window of five turns appears to be an optimal chat length to analyze the context of particular turns in computer-mediated chat during collaborative learning and CPS (Collier, Ruis & Shaffer, 2016; Samei et al., 2014). This amount of context has been explored to improve speech act classification accuracy and to detect multi-turn discourse patterns. Another approach is to construct a STN on adjacent speech acts (Morgan et al., 2012). An STN computes the probabilities of adjacency pairs in a corpus of chat sequences. Stated more formally, it is the transition probability between adjacent speech act categories (SAC) that are indexed by particular team participants: [$\text{P-SAC}_n \rightarrow \text{P-SAC}_{n+1}$]?

Some measures of CPS can theoretically be derived from the distribution P-SAC node categories and the transition probabilities between these node categories. Students who take initiative would have a high proportion of Questions, Requests and Statements, whereas students who are responsive team members (but not leaders) would have a relatively high proportion of Reactions. A disruptive team member would have a high proportion of negative Expressive Evaluations, whereas a social loafer would have a low number of contributions compared with other team members. Regarding the state transitions, responsive team members would have a relatively high transition probability between Questions/Requests of others and the participant's Reactions or Statements; these transition probabilities would be low for unresponsive team members. Thus, these probabilistic metrics have relevance to many of the cells in the CPS assessment matrix. However, available studies have not empirically evaluated the mapping between these automated measures and the 12 cells in the CPS assessment matrix.

Latent Semantic Analyses (LSA) and Semantic Comparisons

LSA (Landauer, Foltz & Laham, 1998) is used to analyze the semantic content of the team members' contributions, a level of language that is not tapped in speech act analyzes. For example, LSA has been used to analyze the coherence of teams and characteristics of individual team members (Dowell, 2017; Foltz & Martin, 2008; Gorman et al., 2003). Text excerpts can be evaluated on semantic similarity through LSA as well as other semantic similarity evaluators that have been described.

Semantic comparison metrics, such as LSA and SEMILAR (Rus et al., 2013), provide an assessment of establishing a shared understanding and building on what each other knows, both of which are theoretically important in the process of establishing and maintaining shared understanding component of collaboration in PISA CPS 2015. The *relevance* (R) of a turn's meaning to the problem being solved is used to compute the extent to which a turn is on versus off topic. This is measured as the semantic overlap between each turn and the semantic topics in the problem being solved. The *givenness* (G) and *newness* (N) of individual turns can be computed for individual team members and the team as a whole (Hempelmann et al., 2005; Hu et al., 2014). A productive collaborative team member contributes relevant information that is new and also builds on other team member's topic-relevant ideas in a responsive fashion. Scores for R, G, and N can be automatically computed by LSA and other semantic evaluators such as SEMILAR, with values that vary from near zero to one. For example, a team member who productively leads the conversation would have a vector of RGN measures such as (0.9, 0.4, 0.6). Team members who echo ideas of others in a conversation would have a (0.9, 0.5, 0.0) vector if they stay on topic, but a (0.0, 0.5, 0.0) vector on off-topic talk. A team member with a (0.0, 0.0, 0.9) vector would be in their own irrelevant worlds and not helpful to collaboration. These profiles have been confirmed in a recent dissertation by Dowell (2017).

There are many other measures of team members and teams that can be computed from these similarity-based metrics and transitions between team members (Dowell, 2017). *Participation* is the relative proportion of a participant's contributions (turns) out of the total number of group contributions; physical actions can be computed in an analogous way to assess the second component of the collaborative dimension (taking appropriate actions to solve the problem) in PISA CPS 2015. *Responsiveness* (analogous to G for givenness) assesses how responsive a team member's contributions are to all other group members' previous contributions in the conversation. *Social impact* measures how turn contributions of a team member have a semantic similarity to other members' contributions in future follow-up responses. *Team member cohesion* measures how semantically similar a team member's contributions are to the same member's previous conversational turns. That is, is a team member saying the same thing over and over? *Communication density* measures how much information in a turn is distinctive to the topic, compared with everyday topics of conversation. All of these metrics can be applied to individual team members as well as the team as a whole.

Epistemic Network Analysis (ENA)

ENA attempts to assess the complex thinking, discourse, reasoning, and topics addressed in professional disciplines and communities (Nash & Shaffer, 2011; Shaffer et al., 2009; Shaffer, Collier & Ruis, in press). There is a disciplinary style of thinking and talking that resonates with the expertise of the community of stakeholders. In scientific disciplines, for example, the discourse might involve claims with supporting empirical evidence and causal analyses. That is a very different discourse from mathematicians or art historians.

ENA's analysis of chat in CPS begins by representing the content as a network structure of connections among critical knowledge, skills, values, and epistemic moves in a professional domain. It measures the strength of association among these cognitive elements and quantifies changes in the composition and strength of those connections over time for individual team members and the entire team. ENA constructs a metric space that enables comparison of individual or group networks through 1) difference graphs, which visualize the differences in weighted connections between two networks, and 2) summary statistics, which reflect the weighted structure of connections in the networks.

It is beyond the scope of this chapter to precisely specify the algorithms that underlie ENA and the process of applying ENA to CPS data (Shaffer, Collier & Ruis, in press, for the ENA Toolkit). The initial step consists of annotating chat turn sequences (i.e., stanzas, sliding turn windows of about length 5) on important cognitive categories (i.e., expressions of skills, knowledge, identity, values, and epistemic content), based on the words expressed in those turns. The next step computes a matrix of co-occurrences of these cognitive categories within these turn sequences and statistically reduces the resulting set of co-occurrence matrices to a small number of dimensions through singular value decomposition. When there are only two or three dimensions, it is possible to plot each cognitive category in a 2- or 3-D metric space; the size of the cognitive category in the space reflects its relative frequency, whereas the thickness of the links between the concept categories reflects the co-occurrence frequency. The resulting network patterns can be compared for different team members, the team as a whole, different phases of CPS interactions, and different chat contexts associated with the profession. ENA has been applied to the land science chat corpora (Collier, Ruis & Shaffer, 2016) and medical engineering design in teams of 3–5 members along with a mentor.

A discipline-oriented style of thinking and talking would of course be an important characteristic to detect and track in team-based ITSs integrated with GIFT. A team member or team as a whole would be regarded as having higher domain expertise to the extent that the chat exhibits higher disciplinary talk that can be automatically quantified from the qualitative input. There are also some links to the CPS assessment matrix of PISA CPS 2015. For example, the discipline thinking parameters are relevant to the problem-solving component D (monitoring and reflecting) and the identity concept categories are relevant to collaboration component 3 (establishing and maintaining team organization). At this point there has been no systematic evaluation of the use of ENA in the scoring of CPS based on the PISA 2015 framework.

Recommendations and Future Research

The most obvious recommendation is to add these automated measures of CPS to GIFT and team-based ITSs. The scores, competencies, and measurements at varying grain sizes would be automatically computed and stored in the Learner Record Store of GIFT. This chapter identifies some automated quantitative algorithms for detecting and assessing many aspects of CPS from the content of the chat logs. These assessments apply to either the entire team or to individual team members as units of analysis. These content-based assessments are a generation beyond the traditional sociometric analyses that compute simple metrics, such as who talks most, who talks to whom, and how many words. Most of the assessments are also aligned with the theoretical framework of a large-scale international assessment, namely PISA CPS 2015. Although this is a promising start, the reliability and validity of these automated assessments await future research.

A second major recommendation is to incorporate these CPS assessments in production rules that formulate what the adaptive, intelligent tutor does next. For example, the suite of applications in GIFT already has the AutoTutor agents that help individuals learn by holding a conversation in natural language (Graesser, 2016; Cai et al., 2015). This could be expanded to include an automated AutoMentor in team contexts when GIFT takes on teams. Some simple production rules were proposed in Graesser et al. (submitted):

1) If the team is stuck and not producing contributions on the relevant topic, then the agent says, "What's the goal here?" or "Let's get back on track."
2) If the team meanders from topic to topic without much coherence, then the agent says, "I'm lost!" or "What are we doing now?"
3) If the team is saying pretty much the same thing over and over, then the agent says, "So what's new?" or "Can we move on?"
4) If a particular team member (Harry) is loafing, the agent says, "What do you think, Harry?"
5) If a particular team member (Sally) is dominating the conversation excessively, the agent says, "I wonder what other people think about this?"
6) If one or more team members express unprofessional language, the agent says, "Let's get serious now. I don't have all day."

Important next steps are to identify a larger set of production rules for CPS, implement them in GIFT environments, and evaluate whether they improve collaborative problem-solving performance (Sottilare et al., 2017).

Acknowledgements

The research on was supported by the National Science Foundation (NSF) (DRK–12-0918409, DRK–12 1418288), the Institute of Education Sciences (IES) (R305C120001), the US Army Research Laboratory (W911INF-12-2-0030), and the Office of Naval Research (N00014-12-C-0643; N00014-16-C-3027). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, IES, or the US Department of Defense. The Tutoring Research Group (TRG) is an interdisciplinary research team comprised of researchers from psychology, computer science, and other departments at University of Memphis (visit http://www.autotutor.org).

References

Cai, Z., Graesser, A.C., Forsyth, C., Burkett, C., Millis, K., Wallace, P., Halpern, D. & Butler, H. (2011, November). Trialog in ARIES: user input assessment in an intelligent tutoring system. In W. Chen & S. Li (Eds.), *Proceedings of the 3rd IEEE International Conference on Intelligent Computing and Intelligent Systems* (pp.429–433). Guangzhou: IEEE Press.

Cai, Z., Graesser, A.C. & Hu, X. (2015). ASAT: AutoTutor script authoring tool. In. R. Sottilare, A.C. Graesser, X. Hu & K. Brawner (Eds.), Design Recommendations for Intelligent Tutoring Systems: Authoring Tools (Vol. 3) (pp.199–210). Orlando, FL: US Army Research Laboratory.

Cannon-Bowers, J. A. & Salas, E. (2001). Reflections on shared cognition. *Journal of Organizational Behavior*, *22*, 195–202.

Clark, H.H. (1996). *Using language*. Cambridge, United Kingdom: Cambridge University Press.

Collier, W., Ruis, A. & Shaffer, D. W. (2016). Local versus global connection making in discourse. In International Conference of the Learning Sciences. Singapore.

Dillenbourg, P. & Traum, D. (2006). Sharing solutions: Persistence and grounding in multi-modal collaborative problem solving. *The Journal of the Learning Sciences, 15*, 121–151.

Dowell, N.M. (2017). A computational linguistics analysis of learners' discourse in computer-mediated group learning environments. Dissertation, University of Memphis.

Dowell, N. M., Graesser, A. C., Cai, Z. (in press). Language and discourse analysis with Coh-Metrix: Applications from educational material to learning environments at scale. Journal of Learning Analytics.

Fiore, S. M., Rosen, M. A., Smith-Jentsch, K. A., Salas, E., Letsky, M. & Warner, N. (2010). Toward an understanding of macrocognition in teams: Predicting processes in complex collaborative contexts. *Human Factors*, *52*, 203–224.

Fiore, S. M., Wiltshire, T. J., Oglesby, J. M., O'Keefe, W. S. & Salas, E. (2014). Complex collaborative problem-solving processes in mission control. *Aviation, space, and environmental medicine, 85*(4), 456–461.

Foltz, P. W., Kintsch, W. & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. Discourse Processes, 25, 285–307.

Foltz, P. W. & Martin, M. J. (2008). Automated communication analysis of teams. In E. Salas, G. F. Goodwin & S. Burke (Eds.), Team effectiveness in complex organisations and systems: Cross-disciplinary perspectives and approaches (pp. 411–431). New York, NY: Routledge.

Funke, J. (2010). Complex problem solving: A case for complex cognition? *Cognitive Processes, 11*, 133–142.

Gorman, J. C., Foltz, P. W. Kiekel, P. A., Martin, M. A. & Cooke, N. J. (2003) Evaluation of Latent Semantic Analysis-based measures of communications content. *In Proceedings of the 47th Annual Human Factors and Ergonomic Society Meeting.*

Graesser, A.C. (2016). Conversations with AutoTutor help students learn. *International Journal of Artificial Intelligence in Education, 26*, 124–132.

Graesser, A.C., Dowell, N.M., Clewley, D. & Shaffer, D.W. (submitted). Agents in collaborative problem solving. *International Journal of Computer Supported Collaborative Learning.*

Graesser, A.C., Forsyth, C.M. & Foltz, P. (2016). Assessing conversation quality, reasoning, and problem solving performance with computer agents. In B. Csapo, J. Funke, and A. Schleicher (Eds.), *On the nature of problem solving: A look behind PISA 2012 problem solving assessment* (pp. 275–297). Heidelberg, Germany: OECD Series.

Graesser, A.C., McNamara, D.S., Cai, Z., Conley, M., Li, H. & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. Elementary School Journal, 115, 210–229.

Graesser, A. C., Penumatsa, P., Ventura, M., Cai, Z. & Hu, X. (2007). Using LSA in AutoTutor: Learning through mixed initiative dialogue in natural language. In T. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.), Handbook of latent semantic analysis (pp. 243–262). Mahwah, NJ: Erlbaum.

Greiff, S., Kretzschmar, A., Müller, J. C., Spinath, B. & Martin, R. (2014). The computer-based assessment of complex problem solving and how it is influenced by students' information and communication technology literacy. *Journal of Educational Psychology, 106*(3), 666–680.

Griffin, P. & Care, E. (2015). ATC21S method. In P. Griffin & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills: Methods and Approach*. Dordrecht: Springer.

Hempelmann, C. F., Dufty, D., McCarthy, P., Graesser, A. C., Cai, Z. & McNamara, D. S. (2005). Using LSA to automatically identify givenness and newness of noun-phrases in written discourse. In B. Bara (Ed.), Proceedings of the 27th Annual Meeting of the Cognitive Science Society (pp. 941–946). Mahwah, NJ: Erlbaum.

Hesse, F., Care, E., Buder, J., Sassenberg, K. & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin and E. Care (eds.), Assessment and teaching of 21st century sills (pp. 37–55). Heidelberg, GA: Springer.

Hu, X., Nye, B. D., Gao, C., Huang, X., Xie, J. & Shubeck, K. (2014). Semantic representation analysis: A general framework for individualized, domain-specific and context-sensitive semantic processing. In D.D. Schmorrow and C.M. Fidopiastis (eds.), *Foundations of augmented cognition: Advancing human performance and decision-making through adaptive systems* (pp. 35–46). Springer International Publishing.

Kuo, B.-C, Liao, C.-H., Pai K.-C., Shih S.-C., Li C.-H. & Mok, M. C. M. (2016). Computer-based collaborative problem solving assessment in Taiwan. Manuscript submitted for publication.

Landauer, T. K, Foltz, P. W. & Laham, D. (1998). An introduction to Latent Semantic Analysis. Discourse Processes, 25, 259–284.

Liu, L., Von Davier, A., Hao, J., Kyllonen, P. & Zapata-Rivera, D. (2015). A tough nut to crack: Measuring collaborative problem solving. In R. Yigal, S. Ferrara & M. Mosharraf (Eds.), Handbook of research on technology tools for real-world skill development (pp. 344–359). Hershey, PA: IGI Global.

Morgan, B., Keshtkar, F., Duan, Y. & Graesser, A.C. (2012). Using state transition networks to analyze multi-party conversations in a serious game. In. S. A. Cerri & B. Clancey (Eds.), Proceedings of the 11th International Conference on Intelligent Tutoring Systems (ITS 2012) (pp. 162–167). Berlin: Springer-Verlag.

Mu, J., Stegmann, K., Mayfield, E., Rosé, C. & Fischer, F. (2012). The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *International Journal of Computer-Supported Collaborative Learning, 7(*2), 285–305.

Nash, P. & Shaffer, D.W.(2013). Epistemic trajectories: Mentoring in a game design practicum. *Instructional Science 41*(4): 745–771

National Research Council (2011). Assessing 21st century skills. Washington, DC: National Academies Press.

OECD (2013). PISA 2015 collaborative problem solving framework. Paris: France: OECD.

OECD (2015). *PISA 2015 released field trial cognitive items.* Paris, France: OECD.

Olney, A., Louwerse, M., Mathews, E., Marineau, J., Hite-Mitchell, H. & Graesser, A. (2003). Utterance classification in AutoTutor. In J. Burstein & C. Leacock (Eds.), Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing. Philadelphia: Association for Computational Linguistics.

O'Neil, H. F., Chuang, S.H. and Baker, E.L. (2010). Computer-based feedback for computer-based collaborative problem-solving. In D. Ifenthaler, P. Pirnay-Dummer, N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 261–279). New York, NY: Springer-Verlag.

Pennebaker, J. W., Booth, R. & Francis, M. (2007). LIWC2007: Linguistic inquiry and word count. Austin, Texas: liwc. net. 2007, Austin, TX. Retrieved from liwc. Net

Rus, V., Lintean, M. C., Banjade, R., Niraula, N. B. & Stefanescu, D. (2013). SEMILAR: The Semantic Similarity Toolkit. In ACL (Conference System Demonstrations) (pp. 163–168).

Rus, V., Lintean, M., Graesser, A. C. & McNamara, D.S. (2012). Text-to-text similarity of statements. In P. McCarthy and C. Boonthum-Denecke (Eds.), Applied natural language processing: Identification, investigation, and resolution (pp. 110–121). Hershey, PA: IGI Global.

Rus, V., Moldovan, Graesser, A.C. & Niraula, N. (2012). Automatic discovery of speech act categories in educational games. In K.Yacef,O. Zaïane, H. Hershkovitz, M. Yudelson, and J. Stamper (Eds.) Proceedings of the 5th International Conference on Educational Data Mining (pp. 25–32). Chania, Greece: International Educational Data Mining Society.

Salas, E., Cooke, N. J. & Rosen, M. A. (2008). On teams, teamwork, and team performance: Discoveries and developments. Human Factors: The Journal of the Human Factors and Ergonomics Society, 50(3), 540–547.

Samei, B., Li, H., Keshtkar, F., Rus, V. & Graesser, A. C. (2014). Context-based speech act classification in intelligent tutoring systems. In S. Trausan-Matu, K. Boyer, M. Crosby & K. Panou (Eds.), Proceedings of the Twelfth International Conference on Intelligent Tutoring Systems (pp. 236–241). Berlin: Springer.

Shaffer, D. W., Collier, W. & Ruis, A. R. (in press). A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics.*

Sottilare, R.A., Burke, C.S., Salas, E., Sinatra, A.M., Johnston, J.H. & Gilbert, S.B. (2017). Towards a Design Process for Adaptive Instruction of Teams: A Meta-Analysis. International Journal of Artificial Intelligence in Education. DOI: 10.1007/s40593-017-0146-z.

Shaffer, D. W., Hatfield, D., Svarovsky, G., Nash, P., Nulty, A., Bagley, E. A., Frank, K., Rupp, A.A., Mislevy, R. J. (2009). Epistemic Network Analysis: A prototype for 21st century assessment of learning. The International Journal of Learning and Media, 1(1), 1–21.

Tausczik, Y. R. & Pennebaker, J. W. (2013). Improving teamwork using real-time language feedback. *Proceedings of Human Factors in Computing Systems (*CHI), 459–468.

Von Davier, A. & Halpin, P. (2013). Collaborative problem solving and the assessment of cognitive skills: psychometric considerations (Research Report No. ETS RR-13-41) (pp. 1–42). Educational Testing Service.

Zapata-Rivera, D., Jackson, G.T. & Katz, I. (2015). Authoring conversation-based assessment scenarios. In R. Sottilare, X. Hu, A. Graesser & K. Brawner (Eds.), Design Recommendations for Adaptive Intelligent Tutoring Systems. (Vol.3)(pp.169–178). Orlando, FL: US Army Research Laboratory.