Collaborative Problem Solving: Innovating Standardized Assessment

Lei Liu, Jiangang Hao, Jessica J. Andrews, Mengxiao Zhu, Robert J. Mislevy, and Patrick Kyllonen
lliu001@ets.org, jhao@ets.org, jandrews@ets.org, mzhu@ets.org, rmislevy@ets.org, pkyllonen@ets.org
Educational Testing Service

Alina A. von Davier, ACT, alina.vondavier@act.org
Deirdre Kerr, Sony Interactive Entertainment, ataride@gmail.com
Thales Ricarte, Universidade de São Paulo, thalesam@icmc.usp.br

Art Graesser (discussant), University of Memphis, art.graesser@gmail.com

Abstract: In this symposium, we present the overall design, data, and scientific findings from the ETS Collaborative Science Assessment Prototype (ECSAP). We are opening our data to the CSCL community and introducing the procedures to request access to the data. ECSAP was developed to explore the assessment of collaborative problem solving (CPS) competency through a large-scale and standardized approach. The goal of this symposium is to examine research questions that are of interest to the CSCL community, such as how CPS skills and collaborative patterns interact with performance outcomes, and how prior content knowledge and personality of team members affect the collaboration process and outcomes. In our study, we collected both individual and collaborative responses (~1500 responses) to the ECSAP instruments. We present our study findings that used new methodologies in psychometrics and followed the best practices of psychometrics and statistics.

Introduction

In contemporary networked and technology-mediated knowledge economies, collaborative problem solving becomes a critical competency for college and career readiness and has been used extensively by educators at all levels. The majority of research on CPS has focused on learning, for example, finding effective ways to promote learning in a (computerized) collaborative environment (Stahl et al., 2006) or developing interventions to foster collaboration skills that contribute to improved learning (Sottilare et al., 2012). However, the assessment aspect of CPS has been relatively less researched. Among the existing studies on assessing CPS, most of them are designed from the perspective of revealing important aspects of CPS (Cohen et al., 1999; DeChurch and Mesmer-Magnus, 2010; O'Neil, 2014; Woolley et al., 2010) based on small samples of participants (von Davier & Halpin, 2013). Many studies collected data from small groups of participants based on convenience, and often did not use standardized assessments in which assessment items, scoring procedures, and interpretations were consistent across test forms. The convenience sampling and the non-standardized instruments in these studies motivate questions about possible bias and the reproducibility of the findings (Hao, Liu, von Davier & Kyllonen, in press).

Among the existing large-scale assessments for CPS, both human-agent (Graesser, Dowell, Clewley, & Shaffer, 2016) and human-human collaborations have been used. In the CPS tasks developed for the *Programme for International Student Assessment* (PISA) in its sixth survey during 2015 (OECD 2013), students collaborated with a different number of virtual partners (agents) on a set of computer-based collaborative tasks and communicated with their virtual partners by choosing from a list of predefined texts. The use of virtual agent and predefined texts is a compromise from a person-to-person collaboration made to ensure standardization, which may pose threats to the validity of assessing collaboration. Another notable assessment for CPS (albeit not standardized) was developed for the Assessment and Teaching of 21st Century Skills project (ATC21S) carried out by Griffin and colleagues (Griffin et al., 2012). In this assessment, two students collaborated via text chat to solve computer-based collaborative tasks. Their actions and response time were automatically coded according to a CPS framework (Adams et al., 2015). Both PISA 2015 and ATC21S consider CPS as a competency that holds across a wide range of domains. In our research, we built off both PISA and ATC21S work and developed ECSAP to explore the assessment of CPS in the domain of science via large-scale data collection and standardized assessment instruments (Liu, Hao, von Davier & Kyllonen, 2015; Hao, Liu, von Davier & Kyllonen, 2015). The ECSAP consists of three assessment instruments to measure the general science knowledge, personality, and CPS. It also includes a background information survey and an after collaboration survey. In this symposium, we bring together a collection of four papers to describe the design of

ECSAP as well as a series of studies to explore several research questions that are of interest to the CSCL field, such as how CPS skills and collaborative patterns interact with collaboration outcomes, and how prior content knowledge and personality of team members affect the collaboration process and outcomes. In presentation 1, we provide an overview of the assessment design and data product from ECSAP. In presentation 2, we introduce a CPS framework that supported the assessment design and discourse analyses. In presentation 3, we present a novel approach for modeling interaction patterns and show how they affect the collaboration outcomes. In presentation 4, we explore how the general science knowledge and team members' personality affect the collaboration outcomes. Our discussant, Dr. Art Graesser, will address how the papers collectively advance CPS assessment in a standardized way and identify gaps in current research and implications for future work.

Presentation 1: ECSAP design and data

Jiangang Hao, Lei Liu, Jessica J. Andrews, Alina A. von Davier, Mengxiao Zhu, and Patrick Kyllonen

The ECSAP was developed to address three major research questions based on large-scale data: identify constructs of CPS and collaboration patterns in the domain of science; find out how the collaborative process affects the collaboration outcome; and explore how the team members' content-relevant knowledge and their personalities affect the collaboration process and outcome. As we are aiming at addressing these questions with large-scale data, we have to compromise on several things to make it practically feasible. For example, to reduce the confounding factors and alleviate the privacy concerns, we chose to use the text-mediated communication rather than video/audio-mediated communication. More details about how we have addressed the practical challenges can be found in Hao, Liu, von Davier, & Kyllonen (in press). There are five instruments in the ECSAP as shown in Figure 1.

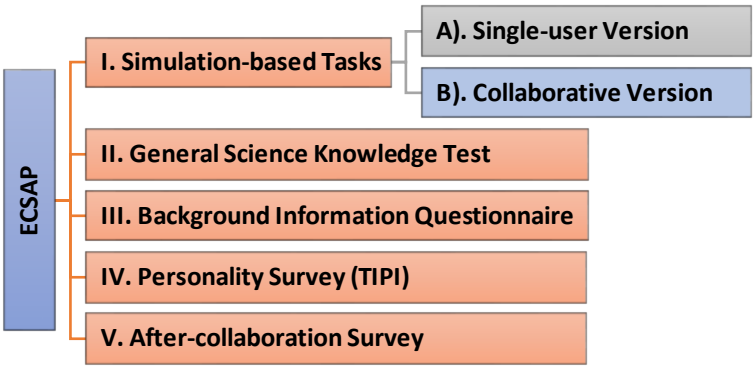

Figure 1. Instruments included in the ECSAP.

The assessment instruments II - V are self-explanatory by name. The simulation-based tasks are modified from an existing game-like science assessment (see details in Zapata-Rivera et al., 2014) in which students work with two virtual agents (a virtual scientist and virtual peer) to solve a complex science problem about making predictions about volcanic eruptions using a dialogue engine. Figure 2 shows the screenshots from the single-user and collaborative versions of the simulation tasks. In the collaborative version of the simulation, students interacted through a chat box. In addition, we designed structured system prompts (based on our CPS framework) to facilitate the collaborative discourse between dyad participants. For each question, we first ask each member of the team to respond the question individually. Then we ask them to collaborate with each other to discuss their answer choices. After collaboration, each member is given a chance to revise his/her initial response. The difference between the initial and revised response captures the gain of the person from the collaboration. Based on the change, we considered the collaboration as effective as long as at least one member made at least a total net change from incorrect to correct. If nobody in the team made at least one total net correct change, we thought of the collaboration as ineffective (Hao, Liu, von Davier, Kyllonen, & Kitchen, 2016). As shown in the second presentation, the effective and ineffective collaboration correspond to different CPS skill profiles.

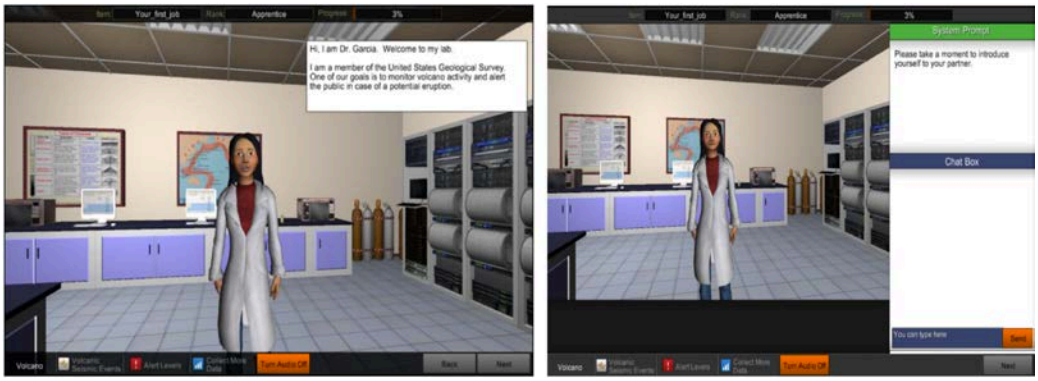

Figure 2. Single-user version (left) and collaborative version (right) of the simulation-based tasks.

We collected the data through Amazon Mechanical Turk, a crowdsourcing data collection platform (Kittur et al., 2008). We recruited 1,500 participants located in United States with at least one year of college education. We administered to them the general science test, personality survey, and demographic survey. Then we randomly selected 500 to complete the single-user version of the simulation. The remaining 1,000 were randomly paired into dyads to complete the collaborative version of the simulation. The data from the simulation task for each team included both the responses to the items in the simulation and the text chat communication between the team members around each item. There were seven multiple-choice-like items in the simulation task, and for each item, there were about five turns of conversation. Seventy-eight percent of the participants were White, 7% were Black or African American, 5% were Asian, 5% were Hispanic or Latino, and 5% were multiracial.

The responses to the multiple-choice-like items (seven such items) were scored based on the corresponding scoring rubrics as presented in Zapata-Rivera et al. (2014). In addition to scoring the outcome responses, we also annotated the chat communication during the collaboration based on our CPS framework (Liu et al., 2015). Two human raters were trained on the CPS framework, and they double-coded a subset of discourse data (15% of the data). The unit of analysis was each turn of a conversation, or each conversational utterance. The raters had two training sessions before they started independent coding. In the first session, the raters were trained on the 33 subcategories of the CPS framework using the skill definitions and coding examples for each subcategory. In the second training session, the trainer and two raters coded data from one dyad together to practice the application of specific codes and address issues specific to classifying utterances using the CPS framework. After the training sessions, the two raters independently coded the discourse data from 79 dyads. One of the 33 subcategories was assigned for each turn, and the inter-rater agreement in terms of unweighted kappa was 0.61 for all 33 subcategories.

Findings based on the aforementioned data and scores/annotations will be presented in the subsequent presentations.

Presentation 2: A CPS framework to support the assessment design and discourse analysis

Lei Liu, Jiangang Hao, Alina A. von Davier, and Patrick Kyllonen

Cognitive and social approaches to science learning have highlighted the importance of collaboration for helping students solve problems and achieve understanding. In educational assessment, there has been a strong recent interest in the evaluation of collaborative problem solving (CPS) as a both a cognitive and social skill (Griffin, Care, & McGaw, 2012; Liu, Hao, von Davier, & Kyllonen, 2015). In our research, we consider collaboration from a discursive perspective as collaboration often takes place in discursive settings (e.g., face-to-face conversations, forum discussion in learning management systems, and chat box in assessment). We define CPS as a process that includes both cognitive and social practices in which two or more peers interact with each other to share and negotiate ideas and prior experiences, jointly regulate and coordinate behaviors and learning activities, and apply social strategies to sustain the interpersonal exchanges to solve a shared problem. This definition describes CPS as both a cognitive and social process (Liu et al., 2015). The cognitive skills include individuals' ability of internalizing others' externalized cognition as well as developing one's own cognition during the problem solving process. The social skills involve individuals' skills of interacting with each other to develop and reach a shared group goal by externalizing one's cognition. In this presentation, we

describe a CPS framework developed based on existing collaborative learning literature, the PISA 2015 CPS Framework (OECD, 2013), and ACTS21 CPS framework (Griffin et al., 2006). There are four major categories in the framework, namely, sharing, negotiating, regulating, and maintaining the communication (see Table 1). Under each major category, there are several subcategories to describe discursive features at the fine grain size (total of 33 subcategories). In addition, we present how we used the CPS framework to analyze dyadic discourse data and how different collaborative patterns emerged and were associated with group performance.

Table 1: CPS framework categories

	Description
Sharing	Conversations about how individual group members bring divergent ideas into a collaborative conversation.
Negotiating	Conversations about the team's collaborative knowledge building and construction through comparing alternative ideas and presenting evidence and rationale of an argument
Regulating	Conversations on clarifying goals, monitoring, evaluating, and confirming the team understanding during problem solving
Maintaining communication	Content irrelevant social communications

As described in Presentation 1, we collected data through a crowdsourcing approach and coded all chat data applying the CPS framework (unweighted kappa was 0.61 for exact match of codes based on 33 subcategories). To highlight collaborative discursive patterns, we introduce a "CPS profile" as a quantitative representation of the CPS skills of each dyad. The profile was defined by the frequency counts of each of the four CPS categories (e.g., sharing ideas, negotiating ideas, regulating problem solving, and maintaining communication). We used unigram and bigram models to represent the CPS profile. The unigram and bigram models are often used in natural language processing to represent text classifications. We adopted similar methods to represent the frequency counts of different CPS skills. We compared the CPS profiles of effective collaboration and ineffective collaboration and found that there were significant differences in the collaborative discourses. When using the unigram models, we found that in general, the effective collaborative teams tended to talk more and they particularly did more discussion to negotiate ideas (see Figure 3). When using the bigram models, we found that the effective teams tended to do more pairs of discussion with negotiations but the ineffective teams tended to do more pairs of discussion with sharing information only (see Figure 3). For example, the bigrams of Negotiate->Share and Negotiate->Negotiate occurred more frequently in effective groups than in ineffective groups. However, the bigram of Share->Share occurred more in the ineffective groups than in the effective groups.

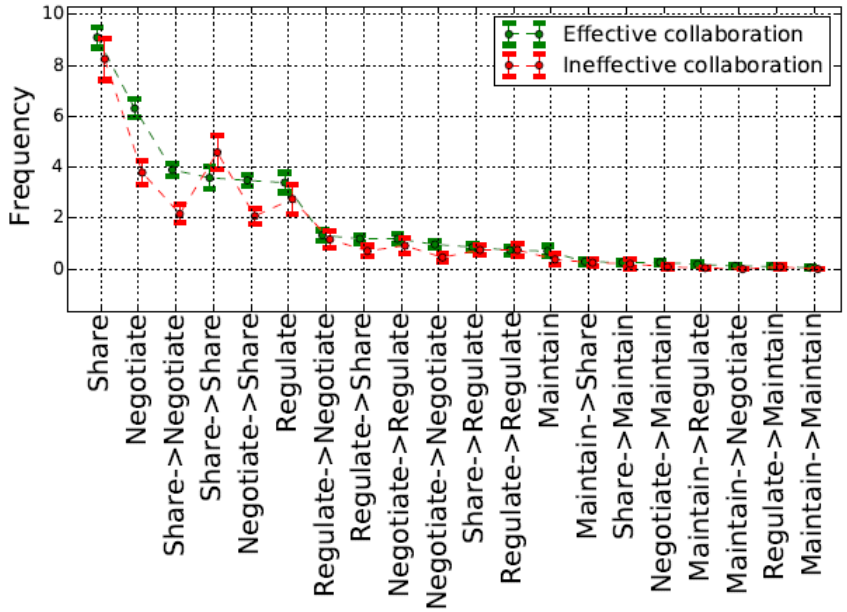

Figure 3. Unigram and bigram profile of CPS skills comparison. The error bars are the standard errors of the means.

Many theoretical and empirical analyses emphasize the importance of active participation and collaboration among students in promoting the effectiveness of online learning. However, there is a need of research to provide empirical evidence of more fine-grained patterns of collaboration that support the effectiveness of learning. Our study attempts to address such a gap. Our findings show that there were differences in dyads' CPS profiles associated with different outcomes of small group collaboration.

Presentation 3: A novel modeling approach for assessing CPS

Jessica J. Andrews, Deirdre Kerr, Robert J. Mislevy, Jiangang Hao, Lei Liu, and Alina A. von Davier

Simulation- and game-based tasks offer opportunities to capture novel sources of assessment data, as these environments afford the capturing of every action taken by students as they engage in game play (Owen, Ramirez, Salmon, & Halverson, 2014). Such affordances particularly lend themselves to capturing evidence of complex skills such as collaborative problem solving (CPS) that are often difficult to measure with more conventional items and tests. One major challenge, however, associated with the use of simulation- and game-based tasks for assessment purposes concerns making sense of the abundance of low-level data generated in these digital environments in order to make claims about individuals or groups. In this paper, we present a novel methodological approach that uses the Andersen/Rasch (A/R) multivariate IRT model (Andersen, 1973, 1995) as an innovative means of modeling interaction patterns. Interaction patterns characterize the ways in which groups interact using log data and performance outcomes. The A/R model addresses tendencies in observations that can be classified into a set of *m* exhaustive and mutually-exclusive nominal categories. In the current instantiation of the A/R model, we model propensities of dyads to behave according to a number of interaction pattern categories. Results from these analyses can be used to answer important questions in collaboration research. We demonstrate specifically how the approach can be used to explore gender and cultural differences in collaborative behavior and how interaction patterns relate to performance outcomes.

The chat logs for each dyad and the log files detailing participants' actions as they completed the ECSAP were coded for interaction patterns that were displayed. The seven items making up the first part of the task were separately coded for interaction patterns (Kappa = .83). That is, patterns were coded at the item level, as each dyad received an interaction pattern code for each of the seven items. Modified versions of Storch (2002) and Tan, Wigglesworth, & Storch (2010) models of dyadic interaction patterns were used to create a rubric for identifying the ways in which participants interacted with their partner. Specifically, the cooperative, collaborative, dominant/dominant, dominant/passive, and expert/novice interaction patterns were included in the rubric. An additional interaction pattern, fake collaboration, was added to the rubric to account for a recurring pattern of behavior not found in the models. Descriptions of each interaction pattern can be found in Table 2.

Table 2: Description of interaction patterns

	Interaction Pattern Description
Cooperative	Both participants share ideas relatively equally, but there is little engagement with the ideas that were shared
Collaborative	Both participants contribute relatively equally and jointly construct a response; engage with each other's ideas; provide explanations and evidence for contributions; critically evaluate other's contributions
Dominant/Dominant	Both participants contribute, but often outwardly seek to maintain own response; difficulty reaching consensus; disagreements
Dominant/Passive	Dominant member takes control of task and shows little effort in inviting contributions from passive member who maintains a passive role
Expert/Novice	One member with more content knowledge (expert) contributes more information, but also encourages contributions from less knowledgeable peer
Fake Collaboration	Both participants contribute information and seem to work together to reach consensus, but revised individual response choices show participants maintained their own (and different) responses

Four of the six interaction patterns (cooperative, collaborative, dominant/dominant, fake collaboration) were displayed as dyads completed the task. The dyad parameter estimates from the A/R model exhibited the propensities for dyads to display each of the coded interaction patterns. In exploring which interaction patterns dyads had the greatest propensity to display, 423 dyads (85.3%) had a tendency to display the cooperative interaction pattern, 62 dyads (12.5%) had a tendency toward the dominant/dominant interaction pattern, 8 dyads (1.6%) had a tendency to toward the collaborative interaction pattern, and 3 dyads (0.6%) had a tendency toward the fake collaboration interaction pattern when compared to all other patterns.

Dyad parameter estimates that corresponded to each of the four interaction patterns were correlated with dyad's team performance. Propensity toward the cooperative (r (494) =.28, $p < .001$) and collaborative (r (494) = .11, $p = .02$) interaction patterns were positively correlated with performance outcomes. Propensity toward the dominant/dominant interaction pattern was negatively correlated with performance outcomes (r (494) = -.21, $p < .001$).

The dyad parameter mean estimates corresponding to each interaction pattern were used to determine whether values were different across same- and mixed-gender and same- and mixed-race dyads. Results revealed no significant differences between same- and mixed-gender and same- and mixed-race dyads in their propensities to display each of the observed interaction patterns.

While there were no significant differences in how the subgroups interacted with each other, we did find differences in how the display of the different interaction patterns related to performance outcomes for the subgroups. For example, comparisons between correlation coefficients showed differences for male-male relative to female-female dyads ($Z = 2.57$, $p = .01$) and male-female dyads ($Z = -2.78$, $p = .01$) for the dominant/dominant interaction pattern. Specifically, the negative correlation between propensity toward the dominant/dominant pattern and performance outcomes was higher for male-male dyads relative to the other subgroups.

These results have important implications for assessment of collaboration, particularly with respect to concerns about fairness and ways of evaluating collaborative behavior. Modeling dyadic interaction patterns using the Andersen/Rasch model is a novel analytical approach for these kind of data and provide an output of a profile for each dyad showing their propensity to behave in accordance with a number of interaction patterns, each of which characterize elements of effective and poor collaborative behavior.

Presentation 4: Relations of individual general science knowledge and personality with collaborative performance

Mengxiao Zhu, Thales Ricarte, Jiangang Hao, Lei Liu, Alina A. von Davier, & Patrick Kyllonen

In collaborative problem solving (CPS), multiple individuals work collectively to solve the problems as a team. As suggested by studies in organization science, team performance can be influenced by many factors both at the individual level, such as individuals' content-related knowledge, and at the team level, such as the team leadership (Mathieu, Maynard, Rapp, & Gilson, 2008). In the assessment of collaborative skills, even though the environment is more confined than in the organizational settings, it is still unclear and remains a very interesting research question how individuals' knowledge/skills and personalities are related to their collective performance as a team. In this study, we focus on two individual level variables, individuals' general science knowledge and personality. The goal is to examine the relationship between these two variables and the individuals' collaborative performance in the task.

This study used the data collected using the ECSAP system introduced in the presentation one of this symposium. Participants' general science knowledge was assessed using a 37-item general science test. To measure individuals' personality, we adopted the Big Five personality traits (McCrae & Costa, 1999). For data collection, to reduce the burden of the participants and at the same time maintain the reliability of the measure, we used the TIPI (Gosling et al., 2003), which contains only ten items and was proven to be a reliable measures of the Big Five personality traits. The analysis included the general science knowledge scores, personality trait measures, as well as scores on the seven items in the simulation-based tasks. For the collaborative version of the simulation task, individuals first worked on the tasks and submitted their individual answers, had a team discussion, and then had the opportunity to revise their final answers. We recorded scores on both the initial scores and the final scores. In many cases, the final scores were better than the individual scores prior to discussion, but there were also cases in which the final scores were lower than the individual scores before discussion.

The reliability of the general science knowledge test is very good, as shown by Cronbach $\alpha = 0.89$ for the 37 general science items. We further categorized the individuals into the low performing and high performing groups using the sample median of 28 as the cutoff value. Individuals who got scores of 28 and

lower were considered as having low general science knowledge (L), and those who got scores higher than 28 were considered as having high general science knowledge (H). For each team, three combinations of individuals were possible, LL, LH and HH. We then compared the performance of these three different types of teams in the CPS tasks. The results showed that, not surprisingly, the HH teams preformed the best for both the initial scores and the final scores, followed by the HL teams and then the LL teams. We also compared the differences between the average initial scores of the two team members and the average final scores to check whether or not different combinations of team members differ in terms of score changes. All three types of teams had positive score increases between the initial scores and the final scores. However, there were no differences in the amount of increase among these different types of teams.

We were also interested in how individuals' general science knowledge would affect their gains from the collaboration. Since each individual can be H or L on general science knowledge, from the individual perspective, there are four types of scenarios, L working with L, L working with H, H working with L, and H working with H. The analysis at the individual level (as shown in Figure 4) showed that everyone benefited from the collaboration regardless of whether it was collaboration of L with H, or H with L, or individuals with equivalent general science skill levels. Among all the possible scenarios, individuals with low general science knowledge benefited the most from the collaboration when working with individuals with high general science knowledge.

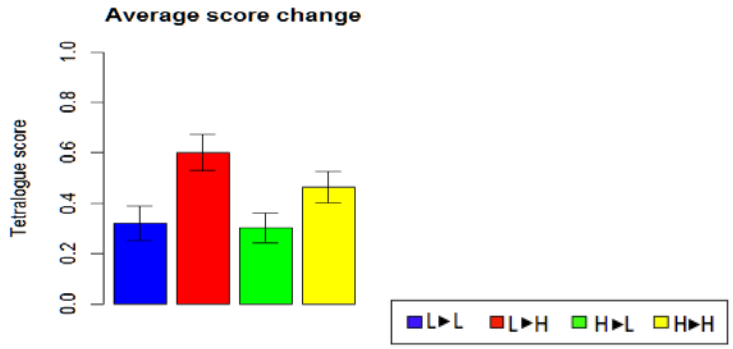

Figure 4. Comparisons among the individuals in terms of the improvement after collaboration. The error bars are the standard errors of the means.

For the personality surveys, since each individual reported on all five personality traits on two items with the 1 to 5 Likert scale, the numbers of teams with different combinations of personality traits are too big, which made it impossible to conduct a similar analysis as in the analysis for the general science skills. Instead, we categorized the teams into three categories based on the collaboration outcomes. During the collaborative process, two team members first submitted their initial answers, and they resubmitted after team discussion. During this process, individuals may or may not change their initial answers, and their scores may also increase, decrease, or stay the same. Based on the individual behaviors and the outcome of the behaviors, we categorized the teams into three groups: Group 0 made no changes and had no score changes; Group 1 made changes and the scores decreased; Group 2 include the teams who made either beneficial changes or harmless changes, that is made changes and scores either increased or stayed the same. We then compared the Big Five personality traits for the three groups. We ran ANOVA analyses and compared the average personality levels for all five dimensions, as well as the absolute difference between the two team members for all five dimensions. The only significant difference, $F(2,\ 382) = 6.23,\ p = 0.002$, was observed for agreeableness, which has characteristics of trust, cooperation, and kindness. Post hoc Tukey tests showed that Group 1 had significantly higher values than Group 0 with a difference of 0.34, and $p = 0.007$; Group 1 also had significantly higher values than Group 2 with a difference of 0.37, and $p = 0.001$. The difference between Group 2 and Group 0 was not significant. These results indicate that high agreeableness was associated with low collaborative performance.

With interesting findings separately on the relations of general science knowledge and personality with collaborative performance, we plan to integrate the studies on these two dimensions and explore the interaction between general science knowledge and personality. During the collaborative processes, we also coded the conversations between the participants, which provide another measure on their collaborative skills. Thus, future work will also examine how the measure of collaborative skills relates to general science knowledge and personality.

References

Andersen, E. B. (1973). *Conditional inference and models for measuring*. Copenhagen: Danish Institute for Mental Health.

Andersen, E. B. (1995). Polytomous Rasch models and their estimation. In G. H. Fischer & I. W. Molenar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 271–291). New York: Springer.

Andrews, J. J., Kerr, D., Mislevy, R. J., von Davier, A. A., Hao, J., & Liu, L. (in press). Modeling collaborative interaction patterns in a simulation-based task. Journal of Educational Measurement.

Dillenbourg, P., & Traum, D. (2006). Sharing solutions: Persistence and grounding in multi-modal collaborative problem solving. *The Journal of the Learning Sciences, 15(1),* 121-151.

Flor, M., Yoon, S., Hao, J., Liu, L., & von Davier, A., (2016), Automated classification of collaborative problem-solving interactions in simulated science tasks, The 11th Workshop on Innovative Use of NLP for Building Educational Applications

Gaudet, A. D., Ramer, L. M., Nakonechny, J., Cragg, J. J., & Ramer, M. S. (2010). Small-group learning in an upper-level university biology class enhances academic performance and student attitudes toward group work. *PLoS ONE*, *5*(12), 1–10.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. Journal of Research in personality, 37(6), 504-528.

Graesser, A., Dowell, N., Clewley, D, & Shaffer, D. (2016), Agents in collaborative problem solving, Manuscript submitted for publication

Griffin, P., Care, E., & McGaw, B. (2012). The changing role of education and schools. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching 21st century skills* (pp. 1-15). Heidelberg, Germany: Springer.

Halpin, P., von Davier, A., Hao, J., & Liu, L. (in press), Measuring student engagement during collaboration, Journal of Educational Measurement

Hao, J., Liu, L., von Davier, A. A., & Kyllonen, P. C. (in press). Initial steps towards a standardized assessment for CPS: Practical challenges and strategies. In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), Innovative Assessment of Collaboration. New York: Springer.

Hao, J., Liu, L., von Davier, A., Kyllonen, P., & Kitchen, C., (2016). Collaborative problem-solving skills versus collaboration outcomes: findings from statistical analysis and data mining, Proceedings of the 9the International Conference on Educational Data Mining.

Hao, J., Liu, L., von Davier, A., & Kyllonen, P. (2015), Assessing collaborative problem solving with simulation based tasks, proceeding of 11th international conference on computer supported collaborative learning, Gothenburg, Sweden

Liu, L., Hao, J., von Davier, A., Kyllonen, P., & Zapata-Rivera, D. (2015). A tough nut to crack: Measuring collaborative problem solving. Y. Rosen, S. Ferrara, & M. Mosharraf (Eds). Handbook of Research on Computational Tools for Real-World Skill Development. Hershey, PA: IGI-Global.

Mathieu, J., Maynard, M. T., Rapp, T. L., & Gilson, L. L. (2008). Team Effectiveness 1997-2007: A Review of Recent Advancements and a Glimpse into the Future. Journal of Management, 34, 410–476. Journal Article.

McCrae, R. R., & Costa Jr, P. T. (1999). A five-factor theory of personality. Handbook of personality: Theory and research, 2, 139-153.

OECD (2013). *PISA 2015 collaborative problem solving framework*. Paris: France: OECD. Retrieved from http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

Owen, V. E., Ramirez, D., Salmon, A., & Halverson, R. (2014, April). *Capturing learner trajectories in educational games through ADAGE (Assessment Data Aggregator for Game Environments): A click-stream data framework for assessment of learning in play*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Rummel, N., & Spada, H. (2005). Learning to collaborate: An instructional approach to promoting problem-solving in computer-mediated settings. *Journal of the Learning Sciences, 14,* 201-241.

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning*, *52*(1), 119–158.

Van den Bossche, P., Gijselaers, W. H., Segers, M., & Kirschner, P. A. (2006). Social and cognitive factors driving teamwork in collaborative learning environments team learning beliefs and behaviors. *Small group research, 37(5),* 490-521.

Zapata-Rivera, D., Liu, L., Chen, L., Hao, J. and von Davier, A.A., (2016). Assessing Science Inquiry Skills in an Immersive, Conversation-Based Scenario. In Big Data and Learning Analytics in Higher Education (pp. 237-252). Switzerland: Springer International Publishing.