Assessing Collaborative Problem Solving with Simulation Based Tasks

Jiangang Hao, Educational Testing Service, jhao@ets.org
Lei Liu, Educational Testing Service, liu001@ets.org
Alina von Davier, Educational Testing Service, avondavier@ets.org
Patrick Kyllonen, Educational Testing Service, pkyllonen@ets.org

Abstract: Assessing collaborative problem solving (CPS) is an integrated part of the computer-supported collaborative learning (CSCL). We present some preliminary results from a project developed for assessing the CPS using web-based simulation. In the simulation, two participants collaborate via a chat box to complete a task on volcano science. By comparing the responses from 486 individuals and 278 teams (dyads) recruited from Amazon Mechanical Turk, we found the performance from the teams (dyads) is significantly higher than that from individuals. We also find that the item difficulty in the simulation affects both the processes and outcomes of the collaboration.

Keywords: collaborative problem solving, simulation, assessment, natural language processing

Introduction

Collaborative problem solving (CPS) is one of the 21st century skills emphasized by the Assessment and Teaching of the 21st Century Skills (ATC21S, 2012). However, developing a psychometrically rigorous assessment for CPS encounters a number of challenges. The construct underlying CPS is very complex, depending on multiple factors, such as the type of the task, the skills and the personalities of the team members, etc. Moreover, CPS has both cognitive aspect and social aspect, and the outcomes from a CPS task are generally the results of the interaction of both. In-depth analysis of the CPS process is probably the only way to disentangle the contribution from social aspect and from cognitive aspect.

In existing literature, many efforts for assessing CPS have focused on the social aspect (de Jong, 2012; DeChurch & Mesmer-Magnus, 2010; Griffin et al., 2012; OECD, 2013). On the other hand, some researchers consider the cognitive aspect of the team as a whole and report the outcomes based on a notion of collective intelligence (O'Neil, 1997; Cohen et al., 1999; Woolley et al., 2010). However, all these researches neglect the collaborative process that is vital for measuring the CPS (von Davier & Halpin, 2013). Von Davier & Halpin (2013) also emphasize the importance to measure individual's cognitive skills in a CPS task and present some psychometric modeling schemes for assessing the process of collaboration.

In this paper, we present our preliminary results from a simulation-based task for assessing CPS. In the task, two human participants work remotely via a chat box to complete a science task about volcano. The goal of this project is to log all the turn-by-turn and time-stamped process data during the collaboration in addition to the outcome data to probe the relations between the cognitive aspect and social aspect of CPS, and between the outcomes and the processes of collaborations with large empirical data.

Methods

Simulation tasks

We developed two simulation tasks for this project, Volcano single and Volcano CPS, based on an earlier simulation (Zapata et al., 2014). The two simulation tasks are almost identical except that the Volcano CPS has an additional chat box layer that allows for collaboration by texting. In Figure 1, we show the screenshots of the two simulations. In the CPS version, for each question/item in the simulation, the two participants are prompted to answer the question by the following four steps: first, each person responds separately. Second, they are prompted to discuss with their partners to get the best answer. Third, after discussion, they are given an opportunity to revise their initial answers. Finally, one of the participants in the team is randomly chosen as team representative to submit a team answer. We record all their responses as well as their conversations in a structured log file (Hao et al., 2015). Such a design permits more effective tracking.

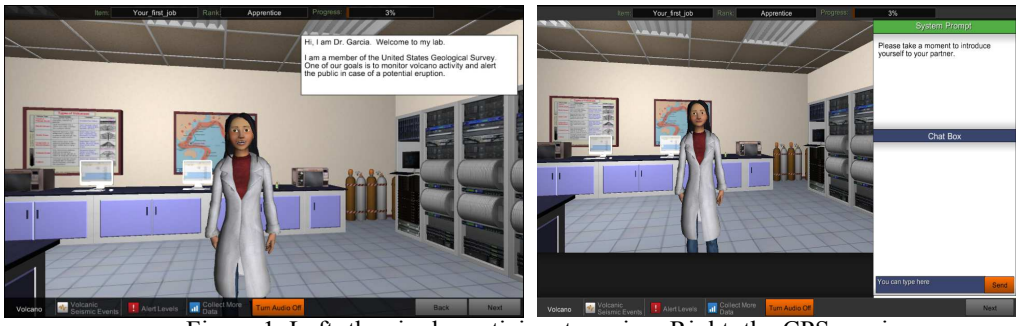

Figure 1. Left: the single participant version. Right: the CPS version.

Participants

Our data collection strategy is crowdsourcing, which has become popular for data-driven research in cognitive and social science (Bridgeman et al, 2012). Amazon Mechanical Turk is a well-developed and supported crowdsourcing web service. For the CPS version, we randomly assign team (dyad) members. That is, when one person started a session, the next person to join is completely random and based on the arrival time. Given the large sample, we can ensure there are adequate numbers of teams in each combination of individual factors.

One limitation of Amazon Mechanical Turk for data collection is that only adults can participate. In our study, we restrict the participants to be college students. We are aiming at collecting data from 500 participants using the single participant version and from 500 dyads using the CPS version. In Table 1, we summarize the participants whose responses are collected so far. As the data collection and analysis are still ongoing, we report only some preliminary results in this paper. It is worth noting that to ensure the samples are equivalent, we have an external test consisting of 37 multiple choice items about general science knowledge to check the equivalence of the groups. The science test has a Cronbach's alpha of 0.89. Each participant needs to take this science test and the mean sum scores and standard deviations of the participants who take the single player version and CPS version are: 26.98 +/- 6.93 for single player group and 26.99 +/- 6.69 for CPS group.

Table 1: Participants summary

Version	Male	Female	Total
Single participant	290	196	486
CPS	280	276	556
Total	**570**	**472**	**1042**

Analysis

Quantitative analysis

All the responses in the simulation are recorded into structured log files. For the conversations, we developed a scoring rubric based on the definition of CPS by PISA 2015 (Graesser, A., & Foltz, P., 2013). The conversations are classified into four categories of social skills by human raters (Liu et al., 2014). At current stage, the scoring is still in progress and we will report the findings in the future. However, we can still find some features from the conversations as "proxies" of the collaboration. For example, the number of words in the conversation could be a good estimate of how enthusiastic the conversations/collaborations are.

On the other hand, each participant answers the science questions in the simulation, which provides information about the collaboration processes and outcomes. In this paper, we will use the sum scores of the first seven selected response (SR) items. It is worth noting that the reliability of seven items is generally not high, about 0.65 in terms of the Cronbach's alpha. We can boost the reliability if we have more items included after we complete the scoring. We performed the following analyses:

First, we check whether the individual initial responses to the items in the CPS version are statistically similar to those in the single-participant version by single participant. If they are similar, that means we can use the individual initial response in the CPS version as a measure of the individual skill. Second, we check whether the revised responses are improved compared to the initial responses. If it is improved, that will be a strong sign that the collaboration does have effects. Third, we check whether the team responses by the team representatives are different from the initial and revised responses. This will inform us any performance changes due to the role change in the team.

In addition to the overall performance, we also want to check the relation between item property and the collaboration processes as well as the collaboration results. In this preliminary analysis, we choose a specific item property, the item difficulty, which is defined as the proportion of correct responses to a given item. The item difficulty parameter is calculated for each item by using the responses from the single participant version. We use the number of words in the conversations for each item as a measure of collaboration process, i.e., how heated the collaboration is. For the collaboration results, we introduce the following response change parameter for each item:

$$\delta = R_1 - I_1 + R_2 - I_2$$

where *R* and *I* refer to the scores of the revised responses and initial responses to that item. The subscripts 1 and 2 refer to the first and second participants in each dyad. If δ is positive, that means there is a positive change in the response after the collaboration and we use δ as a measure of the results of the collaboration. We will check how the mean δ and the mean number of words depend on the item difficulty.

Findings

Table 2 below shows the comparison of the performance in terms of the sum scores of the first seven SR items. Dataset 1 contains the responses from 486 participants who took the single participant version of the simulation. Dataset 2 contains the individual's initial responses from those who took the CPS version. Dataset 3 contains the revised responses from those who took the CPS version and Dataset 4 contains the responses from those who are selected as team representatives to submit team responses.

Table 2: Comparison of the group means of different datasets

Dataset	Mean sum score	Standard error of the mean sum score	Standard deviation of the sum score
1	4.796	0.063	1.389
2	4.842	0.053	1.250
3	5.147	0.049	1.155
4	5.237	0.068	1.134

Based on Table 2, we run two sample t-tests and our findings are as follows. First, the individual initial responses in the CPS version are statistically similar to those from participants who took the single participant version. This means that the individual initial responses in the CPS version can serve as measures of the individual skills. Second, in the CPS version, the revised responses are statistically significantly better than the initial responses, which substantiates that the collaboration does have positive effect on the performance. Third, the team responses by the team representatives are better than the initial responses in a statistically significant way. However, the team responses are not significantly better than the revised responses. This means that the role change of the participants to team representatives does not further change many of their revised responses.

Figure 2 below shows how the collaboration processes and collaboration results (e.g., δ) are dependent on the item difficulty. From the left panel of Figure 2, we observe that the item difficulty does affect the number of communications. For moderate and difficult items (e.g., item 1 to item 4), there are much more communications than that for easy items (e.g., item 5 to item 7). The difference is statistically significant. From the right panel, we observe that the δ is small for very easy items (e.g., item 5 to item 7). This can be understood as follows: for very easy items, their initial responses are almost the best answers and there is little room for improvement after collaboration. Therefore, the δ is small. It is interesting to note that both figures show certain linear relations with the item difficulties, which suggests that the linear relations might be generalizable to other tasks with appropriate changes of the slopes and intercepts.

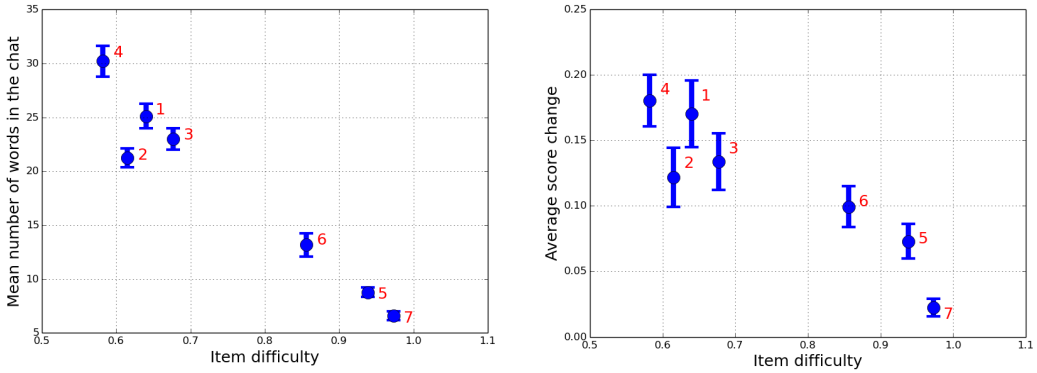

Figure 2. Left: mean number of words in the conversations for each item. The numbers in red fonts are the item number. The blue dots and error bars are the mean and standard errors of the means respectively. Right: mean δ for different items.

Conclusions and implications

In this preliminary analysis, we have shown that there is a significant performance improvement after collaboration and the collaboration processes as well as the collaboration results are dependent on the item difficulty. Though the observed effect size is small, these findings will provide guidance for us to design more appropriate CPS tasks in future. For example, looking at the right panel of Figure 2, if all items are similar to item 1, we can get much bigger increase. That way, we can get a much better effect size in terms of the sum score.

References

De Jong, J. H. A. L. (2012). Framework for PISA 2015: What 15-year-olds should be able to do. Paper presented at the 4th annual conference of the Educational Research Center

Cohen, E. G., Lotan, R. A., Scarloss, B. A., & Arellano, A. R. (1999). Complex instruction: Equity in cooperative learning classrooms. *Theory Into Practice*, 38(2), 80–86.

DeChurch, L. A., & Mesmer-Magnus, J. R. (2010). The cognitive underpinnings of effective teamwork. A meta-analysis. *Journal of Applied Psychology*, 95, 32–53.

Graesser, A., & Foltz, P. (2013). The PISA 2015 Collaborative Problem Solving Framework

Griffin, P., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and teaching of 21st century skills*. New York, NY: Springer.

Hao, J., Smith, L., Mislevy, R., von Davier, A., (2015). *Data model for the log files of game or simulation based assessments*, in preparation

Liu, L., Hao, J., von Davier, A., Kyllonen, P., & Zapata-Rivera, D. (accepted). *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.

O'Neil, H. F. (Ed.). (1997). *Workforce readiness: Competencies and assessment*. Mahwah, NJ: Lawrence Erlbaum Associates

Organization for Economic Co-operation and Development. (2013). *PISA 2015 Draft collaborative problem solving assessment framework.* Paris, France: OECD Publishing.

Von Davier, A., Halpin, P., (2013). Collaborative Problem Solving and the Assessment of Cognitive Skills: Psychometric Considerations, *Research Report*. ETS RR-13-41

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science,* 330(6004), 686–688.

Zapata-Rivera, D., Jackson, T., Liu, L., Bertling, M., Vezzu, M., & Katz, I. R. (2014). Assessing science inquiry skills using trialogues. In S. Trausan-Matu, K. Boyer, M. Crosby & K. Panourgia (Eds.), Intelligent Tutoring Systems (Vol. 8474, pp. 625-626): Springer International Publishing.