International Society of the Learning Sciences

Comparing Collaborative Problem Solving Profiles Derived from Human and Semi-automated Annotation

Jessica Andrews-Todd, Educational Testing Service, jandrewstodd@ets.org
Jonathan Steinberg, Educational Testing Service, jsteinberg@ets.org
Samuel L. Pugh, University of Colorado Boulder, samuel.pugh@colorado.edu
Sidney K. D'Mello, University of Colorado Boulder, sidney.dmello@colorado.edu

Abstract: New challenges in today's world have contributed to increased attention toward evaluating individuals' collaborative problem solving (CPS) skills. One difficulty with this work is identifying evidence of individuals' CPS capabilities, particularly when interacting in digital spaces. Often human-driven approaches are used but are limited in scale. Machine-driven approaches can save time and money, but their reliability relative to human approaches can be a challenge. In the current study, we compare CPS skill profiles derived from human and semi-automated annotation methods across two tasks. Results showed that the same clusters emerged for both tasks and annotation methods, with the annotation methods showing agreement on labeling most students according to the same profile membership. Additionally, validation of cluster results using external survey measures yielded similar results across annotation methods.

Background

Collaborative problem solving (CPS) has received increased attention in the assessment and measurement community over the last several years. This is partly in response to the challenges we face in today's world (e.g., climate change, overpopulation, poverty, and homelessness), which require groups of individuals to come together to solve complex problems (Graesser et al., 2018). There is a growing desire to ensure individuals have the necessary skills to work effectively with others to solve problems and come up with innovative ideas. However, one challenge with assessing individuals' CPS capabilities is identifying targeted skills from their behaviors, particularly when individuals are interacting in digital environments (Andrews-Todd & Forsyth, 2020) where team members are often not in the same physical space.

In many current instances, human-driven approaches (e.g., qualitative coding) are used to identify evidence of individuals' CPS skills, often due to limitations related to resources for and expertise in natural language processing techniques. Machine-driven approaches (e.g., machine learning algorithms) have promise in that they can potentially save time and money with respect to automatically identifying individuals' CPS skills; however, the reliability or comparability of these approaches to human-driven approaches is not well known (Flor et al., 2016; Hao et al., 2017; Pugh et al., 2021; Stewart et al., 2019). In the current study, we compare profiles used to describe learners' CPS behaviors in online collaborative tasks that were generated from human coding and semi-automated (via natural language processing and machine learning) techniques to explore how closely the approaches align. Such an exploration will provide insights into the feasibility of using similar advanced techniques for assessment purposes.

Method

Participants

A total of 88 students in 7th-9th grades from two sites were part of the analysis. The large majority (n=70; 80%) came from a single school site while the rest participated in a lab setting. The individuals completed the study in groups of two (i.e., 44 groups) that were randomly assembled. Of those students who reported their gender/race, 35% were males, 13% were White, 26% were Black or African American, 14% were Asian, 2% were American Indian or Alaska Native, 16% reported being more than one race, 7% reported Other, with 3% unreported. For ethnicity, 19% reported being Hispanic. The average age among students was 13.6 with a range of 12 to 15.

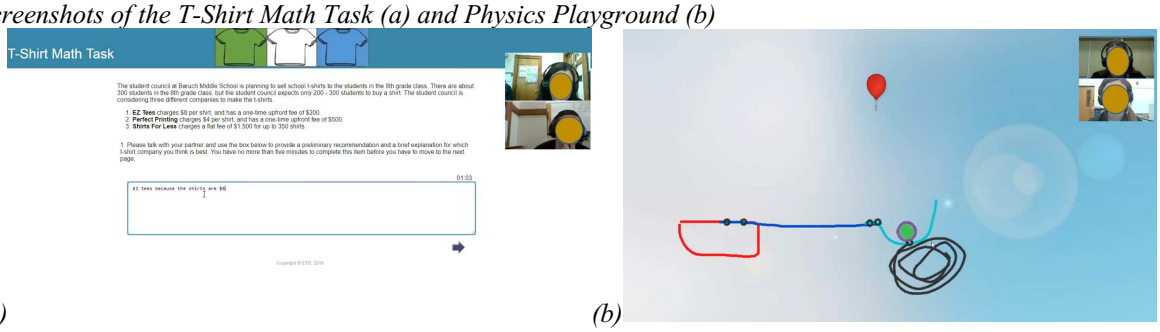
Tasks

The study involved two online tasks, a mathematics task on argumentation and linear functions called the T-Shirt Math Task (Andrews-Todd et al., 2019) and an open-ended physics educational game called Physics Playground (Shute et al., 2013). For the T-Shirt Math Task, students worked together through a series of task items to ultimately determine which of three t-shirt companies was the best choice for purchasing t-shirts for classmates. Only one student could control the cursor at one time, and students alternated control at will. For Physics Playground, students worked together through a series of task levels to draw various objects (e.g., springboard,

ramp, pendulum) to make a ball hit a balloon target. In this task, one student controlled the cursor for three levels (or half the allotted time) and then the other student controlled the cursor for the final three levels. See Figure 1 for screenshots of the tasks.

Figure 1

Screenshots of the T-Shirt Math Task (a) and Physics Playground (b)



(a) *(b)*

Study procedure

Students were randomly assigned to pairs and were seated at individual computer workstations. Students first completed a series of pre-surveys (e.g., content pre-test, background information questionnaire). A researcher then set up Zoom video conferencing software to record students' computer screens, faces, and voices as they worked together to complete the two CPS tasks (the order of the tasks was counterbalanced across teams). After completing the tasks, students completed a series of post-surveys (e.g., content post-test, self-report survey of display of various CPS skills). Only aspects germane to the present study are discussed below.

Human annotation

Video recordings of students' interactions were coded by three raters at the turn level for the presence of one of nine CPS skills across a social (collaboration, teamwork) or cognitive (problem solving, task work) dimension. The social dimension included maintaining communication (content irrelevant social communication), sharing information (content relevant information), negotiating (identifying and resolving conflicts), and establishing shared understanding (communication to learn the perspective of others and ensure statements are understood). The cognitive dimension included exploring and understanding (exploring the environment and building a mental representation of the problem), representing and formulating (representing the problem and generating hypotheses), planning (developing a strategy to solve the problem), executing (actions and communication to carry out the plan), and monitoring (monitoring progress toward the goal and the team organization) (Andrews-Todd & Forsyth, 2020). The raters triple coded 20% of the teams to establish reliability (median ICC across CPS skill ratings was .93), and the remaining data were coded independently.

Semi-automated annotation

We adopted a supervised machine learning approach to automatically classify (annotate) each turn with one of the CPS skills described above. Two skills (exploring and understanding, representing and formulating) were excluded from this analysis due to low prevalence (< 1% of turns). Using human-generated transcripts of each turn, we trained a supervised natural language processing classifier (Bidirectional Encoder Representations from Transformers, or BERT; Devlin et al., 2019) to predict the corresponding CPS skill for that turn. We then summed the predictions to generate estimated frequencies of each skill for both tasks. The models were quite accurate, with correlations between estimated and human-coded frequencies (Spearman rho) ranging from .76 to .93 for Math, and .70 to .95 for Physics. Additional details on classifier design, training, and evaluation are presented in Pugh et al. (2021).

Analyses

Consistent with prior work with a post-secondary student sample interacting with an online collaborative electronics task (Andrews-Todd et al., 2018), we chose an exploratory clustering method (Ward, 1963) for uncovering potential profiles of collaborative problem solvers. The aim in part was to determine how well we would recover a similar number and composition of these profiles with a different population and different tasks. Additionally, the sample size (N=88) did not warrant more robust methods like K-means, typically applied to larger samples. Accordingly, we clustered (with Ward's minimum variance method) the human-coded and semi-automated annotation frequencies of the seven selected CPS skills displayed in each task to allow us to examine the breakdown of possible clusters so that a meaningful number could be chosen.

Results

A three-cluster solution was most defensible from a theoretical perspective and based on the expected relationships to other variables for both human-coded and semi-automated samples and across both task domains. Relative to the total sample, we found one cluster with high amounts of displayed CPS skills which we call Active Collaborators, one with low amounts which we call Social Loafers, and one in the middle of these above two called Middle Performers; the base rates in Table 1 show the sample size percentages for these solutions. Specifically, the learners in the three clusters differed systematically in the average frequencies of CPS skills that were displayed. Though the same clusters emerged for both human-coded and semi-automated samples, there were some differences in how students were categorized with respect to profile membership, as shown in the Table 1 contingency table. Specifically, for Math, 76% of students had the same profile membership across annotation methods, but 24% of students had a different profile membership. For Physics, 70% of students had the same profile membership across annotation methods and 30% had a different profile membership. Agreement (Kappa) for profile membership across annotation methods was .61 for Math and .52 for Physics.

Table 1

Profile Membership According to Human-coded and Semi-automated Methods

	Contingency Table (%)						**Base Rates (%)**			
	Semi-automated (Auto)									
	Math (n = 78)			Physics (n = 70)			Math		Physics	
Human-coded (HC)	Social	Middle	Active	Social	Middle	Active	HC	Auto	HC	Auto
Social Loafers	32.1	16.7	0.0	28.6	1.4	10.0	43.6	35.9	37.1	47.1
Middle Performers	0.0	7.7	0.0	18.6	4.3	0.0	48.7	39.7	40.0	47.1
Active Collaborators	7.7	0.0	35.9	0.0	0.0	37.1	7.7	24.4	22.9	5.7

The key behavioral variables used for validating the respective cluster solutions included performance on the task, performance on a pre-task content knowledge assessment, and the self-report survey (CPS Inventory) in which students reported the extent to which they exhibited CPS behaviors while doing the tasks. Kruskal-Wallis test results revealed patterns were similar across clusters derived from human and semi-automated methods for the math task, with cluster differences emerging for task performance (human-coded: $X^2(2, 78)=11.77$, $p = .003$; semi-automated: $X^2(2, 78)=5.10$, $p = .08$) and the content pre-test (human-coded: $X^2(2, 78)=8.11$, $p = .02$; semi-automated: $X^2(2, 78)=5.34$, $p = .07$), though results only approached significance for the clusters from estimated frequencies. For both measures, Active Collaborators showed higher scores than both Social Loafers and Middle Performers. For the physics task, patterns were also similar across clusters derived from human and semi-automated methods, with cluster differences emerging for the CPS Inventory (human-coded: $X^2(2, 70)=5.80$, $p = .06$; semi-automated: $X^2(2, 70)=5.62$, $p = .06$) and task performance (human-coded: $X^2(2, 70)=15.60$, $p < .001$; semi-automated: $X^2(2, 70)=7.95$, $p = .02$), though the former measure only approached significance. For both measures, Social Loafers performed worse than the other two profile groups.

Discussion

We examined CPS skill profiles derived from human and semi-automated annotation methods to explore the extent to which cluster solutions are comparable across annotation methods. Results showed that similar three-cluster solutions emerged across both annotation methods and across both the T-Shirt Math Task and Physics Playground. Specifically, there were Active Collaborators who on average displayed higher amounts of CPS behaviors relative to other groups, Social Loafers who tended to display few CPS behaviors relative to other groups, and Middle Performers who tended to display CPS behaviors at rates in between the previous two groups. We found that across both tasks, there was general agreement between human-coded and semi-automated methods on how to categorize most students. Interestingly, for instances in which profile membership differed across the annotation methods, patterns of disagreement were different for the math and physics tasks. For the math task, students tended to be in lower performing profiles for the human-coded sample relative to the semi-automated sample. The opposite was true for the physics task where most students were in higher performing profiles for the human-coded sample relative to the semi-automated sample. We also found that patterns of results were similar across annotation methods for validation of clusters using external surveys.

Study results provide preliminary evidence that the current semi-automated annotation method (BERT) shows some promise in providing results comparable to human-coded methods, which we consider the ground truth. This is important when considering approaches that can support efforts to scale up CPS assessment in ways that allow for the full scope of CPS to be measured. In particular, some recent large scale CPS assessment work

does not support open communication (OECD, 2013), and for instances in which it does, the content of the communication is limited to chat and sometimes not evaluated (Adams et al., 2015).

One limitation with the current study is that human transcripts were used for the semi-automated method instead of automated transcripts due to relatively low accuracy rates with automatic speech recognition in noisy classroom settings. A second limitation is that our methods were utilized with a relatively small sample and primarily in the context of intermediate and secondary education classrooms. Future work will need to explore these machine learning approaches with a larger sample and with different populations to determine if similar patterns of results emerge. Future studies can also explore alternative automated approaches to determine the extent to which they yield even more comparable results to profiles derived from human-coded methods. Machine learning approaches like the one used in the current study offer promise in lessening the load of resource-intensive human-coding methods to open many possibilities for relevant CPS work.

References

Adams, R., Vista, A., Scoular, C., Awwal, N., Griffin, P., & Care, E. (2015). Automatic coding procedures for collaborative problem solving. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 115–132). Springer.

Andrews-Todd, J., & Forsyth, C. M. (2020). Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Computers in Human Behavior*, *104*, 105759. https://doi.org/10.1016/j.chb.2018.10.025

Andrews-Todd, J., Forsyth, C. M., Steinberg, J., & Rupp, A. A. (2018). Identifying profiles of collaborative problem solvers in an online electronics environment. In K. E. Boyer & M. Yudelson (Eds.), *Proceedings of the 11th International Conference on Educational Data Mining* (pp. 239–245). International Educational Data Mining Society.

Andrews-Todd, J., Jackson, G. T., & Kurzum, C. (2019). *Collaborative problem solving assessment in an online mathematics task* (Research Report RR-19-24). Educational Testing Service.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the North American Association for Computational Linguistics (NAACL)*, 4171–4186.

Flor, M., Yoon, S.-Y., Hao, J., Liu, L., & von Davier, A. (2016). Automated classification of collaborative problem solving interactions in simulated science tasks. In J. Tetreault, J. Burstein, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the 11th workshop on innovative use of NLP for building educational applications* (pp. 31–41). Association for Computational Linguistics.

Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the science of collaborative problem solving. *Psychological Science in the Public Interest*, *19*(2), 59–92.

Hao, J., Chen, L., Flor, M., Liu, L., & von Davier, A. A. (2017). *CPS-Rater: Automated sequential annotation for conversations in collaborative problem-solving activities* (RR-17-58; pp. 1–9). Educational Testing Service. https://doi.org/10 .1002/ets2.12184

OECD. (2013). *PISA 2015 collaborative problem solving framework*. OECD Publishing.

Pugh, S. L., Subburaj, S. K., Rao, A. R., Stewart, A. E., Andrews-Todd, J., & D'Mello, S. K. (2021). Say What? Automatic modeling of collaborative problem solving skills from student speech in the wild. In I.-H. Hsiao, S. Sahebi, F. Bouchet, & J.-J. Vie (Eds.), *Proceedings of the 14th International Conference on Educational Data Mining* (pp. 55–67). International Educational Data Mining Society.

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in Newton's Playground. *The Journal of Educational Research*, *106*, 423–430.

Stewart, A. E., Vrzakova, H., Sun, C., Yonehiro, J., Stone, C. A., Duran, N. D., Shute, V., & D'Mello, S. K. (2019). I say, you say, we say: Using spoken language to model socio-cognitive processes during computer-supported collaborative problem solving. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–19.

Ward Jr., J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244.

Acknowledgements

This material is based upon work supported by the Institute of Education Sciences under Grant R305A170432 awarded to the first author and in collaboration with the University of Colorado Boulder and CRESST and the National Science Foundation (DUE 1745442/1660877). The opinions expressed are those of the authors and do not necessarily represent the views of IES and NSF.