Psychometric Considerations and a General Scoring Strategy for Assessments of Collaborative Problem Solving

ETS RR–19-41

Jiangang Hao
Lei Liu
Patrick Kyllonen
Michael Flor
Alina A. von Davier

December 2019

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Psychometric Considerations and a General Scoring Strategy for Assessments of Collaborative Problem Solving

Jiangang Hao,[1] Lei Liu,[1] Patrick Kyllonen,[1] Michael Flor,[1] & Alina A. von Davier[2]

1 Educational Testing Service, Princeton, NJ
2 ACTNext, ACT Inc., Iowa City, IA

Collaborative problem solving (CPS) is an important 21st-century skill that is crucial for both career and academic success. However, developing a large-scale and standardized assessment of CPS that can be administered on a regular basis is very challenging. In this report, we introduce a set of psychometric considerations and a general scoring strategy around assessing CPS, summarized based on the results of the extensive empirical studies we conducted at Educational Testing Service (ETS) over the past 6 years. Using the ETS Collaborative Science Assessment Prototype as an example, we show how these psychometric considerations have been incorporated into the development of the assessment prototype and how the scoring strategy has been implemented.

Keywords Collaborative problem solving; psychometrics; assessment design; scoring

doi:10.1002/ets2.12276

Collaborative problem solving (CPS) is widely accepted as an important 21st-century skill (Fiore et al., 2017; Griffin, McGaw, & Care, 2012; Organisation for Economic Co-operation and Development, 2013; Roschelle & Teasley, 1995; World Economic Forum, 2015). Assessments of CPS play important roles in teaching and learning CPS, through which CPS can be more accurately defined and improvements can be quantitatively measured. Developing a large-scale and standardized assessment for CPS and administering it on a regular basis is extremely challenging in practice (Hao, Liu, von Davier, & Kyllonen, 2017). Given that the constructs of CPS are very complex and multidimensional, most of the practical challenges in assessing them are a direct or indirect result of a lack of balance between the ambitious goals and the affordability of a realistic assessment program. On one hand, people intend to measure CPS as comprehensively as possible, while on the other hand, developing such tasks requires tremendous synergy across a wide range of disciplines (von Davier, Hao, Liu, & Kyllonen, 2017), and the constraints of time, budget, and psychometric quality always require compromises to be made. A healthy interplay between these two factors is crucial for developing realistic assessments of CPS at scale.

CPS exhibits itself only in complicated collaborative settings, either face-to-face or mediated by computers and the Internet. For a large-scale assessment, developing and administering face-to-face collaborative activities for people to display their CPS skills are prohibitively expensive and time consuming. Advances in computer and Internet technology have made it possible to create virtual collaborative environments, providing a feasible way to carry out collaboration at large scale. Over the last decade, leveraging virtual collaborative environments, several large-scale assessments of CPS have been developed, and reviews of these assessments can be found in von Davier, Zhu, and Kyllonen (2017), Fiore et al. (2017), and Dingler, von Davier, and Hao (2017). Among these attempts, Assessment and Teaching of 21st-Century Skills (ATC21S; Griffin et al., 2012) and the 2015 Program for International Student Assessment (PISA2015; Graesser et al., 2018; Organisation for Economic Co-operation and Development, 2013) are two good examples that have been carried out at large scale. Both assessments target domain-generic CPS constructs defined based on comprehensive literature reviews and use simulation-based tasks to measure them. In the CPS assessment from ATC21S, two students are assigned to a team to complete several tasks collaboratively via text chats. Each student's CPS is scored primarily based on his or her responses and actions in the tasks (Hesse, Care, Buder, Sassenberg, & Griffin, 2015; Scoular, Care, & Awwal, 2017). In the CPS assessment from PISA 2015, each student is placed onto a team with varied numbers of virtual partners who are programmed to provide a "standardized" collaborative environment. The virtual partners are powered by rules specified by the experts, and the students communicate with the virtual partners by choosing predefined text responses. Students'

Corresponding author: J. Hao, E-mail: jhao@ets.org

CPS is measured by their choice of the predefined texts for communication and by their responses in the collaborative tasks (Graesser et al., 2017; He, Davier, Greiff, Steinhauer, & Borysewicz, 2017).

Though ATC21S and PISA 2015 have made groundbreaking progress toward assessing CPS at scale, they all made compromises due to the limitation of technology, the requirements of psychometric quality, and the constraints of budgets and time. For example, in ATC21S, the direct communications between students have not been scored, though they contain rich information regarding students' negotiation skills as targeted by the assessment. The interdependency of the partners' task-relevant proficiency and CPS has not been seriously considered in the scoring model either. In PISA 2015, the communications with virtual partners are through predefined texts set by experts, which may be entirely different from those in real collaboration. Both assessments use simulation-based tasks whose development becomes a significant practical challenge if new tasks need to be developed and administered regularly, as in most of the large-scale testing programs, such as the *SAT*® test and the *GRE*® tests.

Since 2013, Educational Testing Service (ETS) researchers have carried out a series of empirical studies to explore the possibility of a large-scale and standardized assessment of CPS that can be administered on a regular basis. Four main lines of research are being conducted: first, development and refinement of constructs of CPS (e.g., Andrews et al., 2017; Andrews-Todd & Kerr, 2019; Liu, Hao, von Davier, Kyllonen, & Zapata-Rivera, 2015); second, development of technological infrastructure, such as the ETS Platform for Collaborative Assessment and Learning (EPCAL; Hao, Liu, von Davier, Lederer, et al., 2017), the automated annotation systems (e.g., Flor, Yoon, Hao, Liu, & Davier, 2016; Hao, Chen, Flor, Liu, & von Davier, 2017), and the data analytics system (e.g., Hao, Smith, Mislevy, von Davier, & Bauer, 2016); third, development of assessment prototypes in various domains (e.g., Andrews-Todd, Jackson, & Kurzum, 2019; Hao, Liu, Davier, & Kyllonen, 2015; Martin-Raugh et al., in press); and fourth, development of psychometric methodologies (e.g., Halpin, von Davier, Hao, & Liu, 2017; Hao, Liu, von Davier, Kyllonen, & Kitchen, 2016; Hao & Mislevy, 2019; Zhu & Andrews-Todd, 2019; Zhu & Zhang, 2017). While carrying out this research, we developed a set of psychometric considerations for the assessment and task designs as well as a general scoring strategy for assessing CPS. The focus of the current report is to introduce these considerations and strategy and use the ETS Collaborative Science Assessment Prototype (ECSAP) as an example to illustrate how they were implemented in practice.

This report is organized as follows. In the following section, we introduce a set of psychometric considerations for designing CPS assessments. In the third section, we describe a general scoring strategy for assessing CPS and show how statistical and natural language processing (NLP) techniques could be used to automate the assessment. In the fourth section, we demonstrate how the ideas discussed in the second and third sections have been applied to ECSAP. Finally, in the concluding section, we discuss the implications of the current work and its limitations. We also discuss how our ongoing and future work will address these limitations.

Psychometric Considerations of Assessment and Task Design

Psychometrics is far beyond analyzing data post factum using statistical techniques. The psychometric principles, such as validity, reliability, comparability, and fairness (Messick, 1992), need to be carefully weighed to guide the assessment and task design to ensure that the desired psychometric quality can eventually be achieved. CPS is the interactive result of both a cognitive component (problem solving) and a social component (collaboration). The social component is generally more aligned with "will do" than "can do," making it less likely to be reliably measured in a single snapshot assessment. As such, we need to think carefully not only what assessment tasks need to be used (e.g., task designs) but also how to administer these tasks (assessment designs) to achieve the desired assessment goals.

Let us start with some psychometric considerations around the assessment designs. Obviously, to measure CPS, the prerequisite is that people act in a collaborative environment. This means that each person's performance is dependent not only on himself or herself but also on the partners with whom he or she interacts in the collaboration. Therefore, if we want to report an individual's CPS, we need to design the assessment in a way that "marginalizes" the effects due to the partners. In practice, we can achieve this through two types of designs (Hao, Liu, von Davier, & Kyllonen, 2017). The first is a multiteam round robin-like design, where an individual participates in many (parallel) collaborative tasks with properly sampled human partners in each of the tasks. The second is a virtual-partner design, where an individual collaborates with "standardized" virtual partners (Graesser et al., 2017; Rosen, 2018). For the multiteam round robin-like design, the real challenge is how to carry out multiple sessions of assessment with sufficient participants. One possible real-world scenario could be in a school environment, where a student collaborates with one or more classmates who are

randomly assigned every week or month throughout a semester.[1] For the virtual-partner design, a big challenge is that the current AI technology is still far from supporting an open collaboration with humans, though this might be possible under certain highly constrained circumstances. The pros and cons of using virtual agents for assessing collaboration have been widely discussed (Graesser et al., 2017; Herborn, Stadler, Mustafić, & Greiff, in press; Rosen, 2014, 2015).

In addition to being reported at an individual level, CPS can be reported at a group level. For example, it may be interesting to know the average level of CPS of the employees in Company A compared to those in Company B or of the students in State X compared to those in State Y. The stakeholders of the group-level reporting are usually not persons but rather organizations or policy-making agencies. The group-level reporting simplifies the assessment design of CPS considerably. Generally speaking, as long as the teams are formed randomly[2] from different large groups, the average CPS performance is readily comparable.

After introducing the possible assessment design choices for different reporting goals, let us look at the task design through a psychometric lens. Choosing meaningful constructs is the prerequisite step for developing any serious assessment tasks. However, in practice, a meaningful construct does not necessarily guarantee that it can be measured with needed psychometric quality, particularly through a snapshot test that lasts a limited period. Evidence-centered design (ECD; Mislevy & Riconscente, 2006) provides a general framework for designing complex assessment tasks to keep the evidence aligned with the targeted constructs. ECD principles stipulate that any claims about the measurement of the constructs need to get sufficient support from the relevant evidence revealed through the interaction between the test takers and the test instruments. As such, to ensure that a construct can be measured appropriately, the task needs to be so designed that abundant evidence can be readily elicited from the interactions, preferably at a relatively low cost. In the case of CPS, possible evidence of the CPS constructs can be extracted from the actions of the team members during the collaboration process, the collaboration outcomes, and the communications among the team members. Focusing on different evidence sources often leads to different task design considerations.

If one wants to identify evidence from the actions, significant efforts are needed in the task design to ensure that the task can elicit relevant actions and each team member can have an "equal" opportunity to act to exhibit his or her skills. Moreover, the mapping between the actions and CPS constructs is subject to different tasks. Given these requirements, the design used in one task is usually not directly generalizable to other tasks, which could increase the development cost significantly. On the other hand, identifying evidence from the communication data is less demanding, and this evidence is usually easier to generalize across tasks. In particular, Internet-based technology has made computer-supported collaborations widespread in academia and in the workplace (Stahl, Koschmann, & Suthers, 2006), which could generate a great deal of communication among team members in online collaboration. The significant real-world cases suggest that even assessing only computer-supported CPS is of great interest and meaningful, though we must caution that the CPS exhibited through online collaboration may not entirely overlap with CPS exhibited in a face-to-face collaboration (Suthers, Hundhausen, & Girardeau, 2003; Tutty & Klein, 2008; Warkentin, Sayeed, & Hightower, 1997).

In summary, the aforementioned psychometric considerations in the early design stage of a CPS assessment can provide important guidelines for choosing appropriate assessment and task designs, especially if the assessment is intended to be administered at large scale and on a regular basis.

A General Scoring Strategy

When thinking of assessing CPS, one may immediately jump to the idea that CPS can be scored from low to high on a certain scale based on specific scoring rules specified by experts. However, the constructs of CPS are multidimensional in nature and may not be simply mapped to a unidimensional scale based on certain additivity assumptions. In ATC21S and PISA 2015, though the construct frameworks are specified as multidimensional, the CPS is eventually mapped onto a unidimensional scale in their item response theory modeling procedures. A unidimensional scale oversimplifies the rich information from CPS tasks and could potentially undermine the validity of the assessment. As such, we propose the following scoring strategy.

This scoring strategy includes three key components: measures of the collaboration outcomes, representations of the collaboration processes, and a mapping between the process representations and the outcome measures. The outcome measures are usually straightforward to determine for a given CPS task, as they can always be classified into a problem-solved or problem-not-solved category. However, the collaboration process is often less straightforward to classify as high or low owing to its multidimensional nature. What one often observes from a collaboration process is whether or how

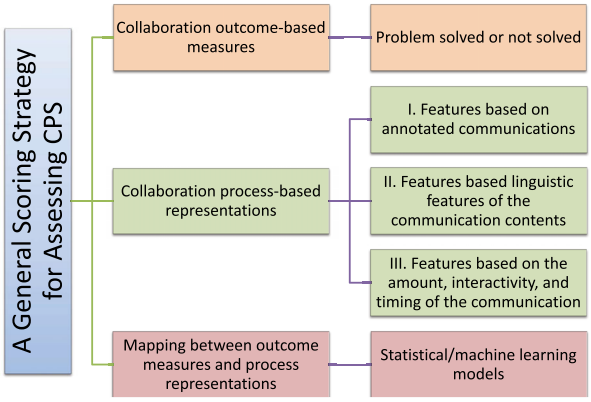

Figure 1 A general scoring strategy for assessing collaborative problem solving.

many of certain skills (e.g., negotiation or sharing) and patterns (e.g., turn taking) were exhibited. These skills and patterns characterize different aspects of the collaboration process and form a feature representation of the process data. In principle, one can directly relate these features to the targeted CPS constructs through an ECD process that is primarily driven by experts' judgments. However, in practice, this is not a straightforward process, as the occurrence of these features does not necessarily relate to the CPS constructs through a simple linear relationship. As such, even experts need more objective criteria to guide the development of the relationship. Among all possible criteria for evaluating CPS, the collaboration outcomes turn out to be simple, objective, and important measures. By mapping the process feature variables to the outcome measures, one can identify which features contribute to the success of the collaboration and which do not. Meanwhile, establishing such a mapping also allows us to develop a real-time intervention mechanism, such as an intelligent facilitator (Hao, Liu, von Davier, Lederer, et al., 2017), for online collaborative learning activities. Figure 1 is a schematic of the scoring strategy.

The central part of this scoring strategy is to develop feature representations for the CPS process. There are many ways to create feature variables from the collaboration process data, depending on the means of communication in the collaboration. If the communication in the collaboration is text mediated, roughly speaking, the features can be categorized into three sets.[3] The first set is based on the annotation of the communications using various frameworks or coding rubrics, which have been widely used in research in computer-supported collaborative learning (CSCL; Andrews et al., 2017; Dönmez, Rosé, Stegmann, Weinberger, & Fischer, 2005; Häkkinen, Järvelä, & Mäkitalo, 2003; Jeong & Hmelo-Silver, 2010; Liu et al., 2015; Strijbos, Martens, Prins, & Jochems, 2006; Weinberger & Fischer, 2006). On the basis of the coded categories, we can readily create numerical feature variables, such as the relative fraction of each category (e.g., unigrams) and consecutive category pairs (e.g., bigrams).

The second set is based on the linguistic features of the communication contents. These features have been extensively explored in the literature on discourse analysis (e.g., Dowell, Graesser, & Cai, 2016; Graesser et al., 2014; Rus, Moldovan, Niraula, & Graesser, 2012). Typical features in this set include sentiment, mood, word complexity, grammatical errors, and so on. The third set is based on the amount, interactivity, and timing of the communication during collaboration. Features in this set include the number of turn takings, response delay, communication density, and statistical measures based on the point process modeling of the timing information (e.g., Halpin et al., 2017; von Davier & Halpin, 2013). We list some specific examples of features in Table 1 but remind readers that this list is expandable.

Generally speaking, the generation of the annotation-based features (Set I) is the most time consuming and laborious, as human coders are needed to code the communications. The advance of NLP technology has made it possible to automate the annotation process once a sufficient amount of human-coded training data are available. For example, we developed an automated annotation system, CPS-rater, by leveraging the interdependency among the turns of the communication data (Hao, Chen, et al., 2017). Such a system outperforms the algorithms that treat each turn as an independent utterance (Flor et al., 2016; Rosé et al., 2008). Most Set II and III features can be generated automatically, as long as the interaction data are adequately logged.

For communications mediated by audio and video, a preprocessing step to transcribe the audio and video is needed before one can generate the three sets of features. In particular, AI-driven technology, such as Amazon Alexa and Google

Table 1 A Partial List of Feature Representations of Collaboration Process

Feature set	Feature name	Details
I	category_i	Proportion of the i*th* annotated category (aka unigram)
	category_i_j	Proportion of the transition from category *i* to *j* (aka bigram)
II	pos_senti	Proportion of positive sentiment
	neg_senti	Proportion of negative sentiment
	neu_senti	Proportion of neutral sentiment
	indicative	Proportion of turns with indicative mood
	imperative	Proportion of turns with imperative mood
	conditional	Proportion of turns with conditional mood
	subjunctive	Proportion of turns with subjective mood
III	n_word	Total number of words in the communication
	n_turn	total number of turns in the communication
	mean_words_turn	Average number of words per turn
	mean_words_minute	Average number of words per minute
	mean_turn_minute	Average number of turns per minute
	min_reply_gap_time	Minimum of the gaps in seconds before replying to partner
	q25_reply_gap_time	25% quantile of the gaps in seconds before replying to partner
	q50_reply_gap_time	50% quantile of the gaps in seconds before replying to partner
	q75_reply_gap_time	25% quantile of the gaps in seconds before replying to partner

Note. We intended to show some examples of the possible features rather than an exhaustive list of the features.

Home, has enabled automated transcription of audio with acceptable accuracy. Meanwhile, the video/audio communication allows the capture of other information, such as tones and affects, which can be turned into feature variables and appended to the three sets of features.

The features discussed are generic across CPS activities and can be computed for each team or each team member. In particular, for studies that target group-level reporting, one can group the teams based on criteria of interest (e.g., good collaboration outcomes vs. bad outcomes, Company A vs. Company B, State X vs. State Y) to examine how the averages of the features in each group are different from other groups. The mapping between the collaboration outcome measures and the process representation is straightforward to obtain by considering the features as independent variables and the outcome measures as dependent variables. Statistical or machine learning models can be used to instantiate the mapping.

This scoring strategy is no longer holding the delivery of a unidimensional score for CPS as its primary goal, though a score can be created based on a certain combination of the process representations and outcome measures. In fact, expecting a unidimensional score from the multidimensional CPS constructs is probably not appropriate at all. This scoring strategy provides more information about the CPS than a unidimensional score can provide. It provides a multidimensional profile based on various aspects of the CPS process and establishes an objective relationship between the collaboration process and collaboration outcomes, which could lead to the development of actionable feedback to participants and an intervention or facilitation mechanism for scaffolding collaborations. We illustrate how the scoring strategy works through an empirical example in the next section.

An Empirical Example

ETS Collaborative Science Assessment Prototype

ECSAP is a project developed to address three main research questions based on large-scale empirical data: (a) identifying suitable subsets of CPS constructs for a standardized assessment at large scale and a regular basis, (b) developing a methodology to evaluate the collaboration process and outcomes, and (c) exploring how team members' task-relevant attributes affect the collaboration (Liu et al., 2017). The comprehensive findings from the ECSAP are beyond the scope of the current report. Instead, we focus on illustrating how the psychometric considerations and the scoring strategy introduced in the previous section have been implemented in ECSAP, which is the first in a series of our research efforts toward assessing CPS. Given the exploratory nature of ECSAP, we decided to limit the assessment to group-level reporting, which greatly simplified the assessment design. We further narrow down our scope by considering only computer-supported collaboration between two people in the domain of science. Finally, we focus on a subset of CPS constructs whose evidence can be extracted from the communications (text chats) during the collaboration.

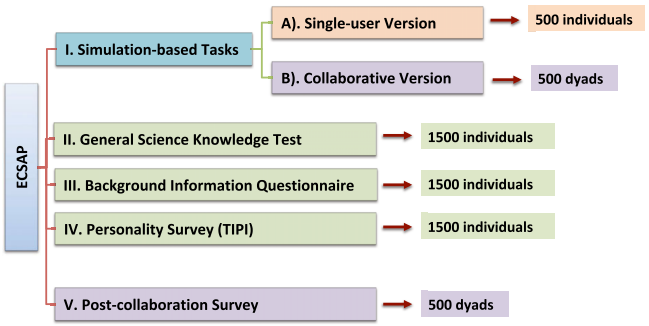

Figure 2 Assessment instruments used in the ETS Collaborative Science Assessment Prototype.

On the basis of the CSCL literature and the assessment frameworks from PISA 2015 and ATC21S, 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" (Liu et al., 2015, p. 344–359, and references therein). Specifically, four CPS constructs—sharing ideas, negotiating ideas, regulating problem solving, and maintaining communication—have been identified as relevant to the CPS activity we are targeting. In addition to the CPS constructs, we identified two potentially task-relevant attributes of each participant: general science knowledge measured with a stand-alone science knowledge test (Rundgren, Rundgren, Tseng, Lin, & Chang, 2012) and the big five personality traits measured with the Ten Item Personality Measure, or TIPI (Gosling, Rentfrow, & Swann, 2003).

In Figure 2, we show the assessment instruments in ECSAP as well as the data we targeted to collect using each instrument. The instruments II–V are self-explanatory by their names. In the single-user version of the simulation-based tasks, one human participant interacts with two virtual agents to solve a complex science problem (Zapata-Rivera et al., 2014). In the collaborative version of the simulation (see Figure 3; Hao et al., 2015), we added a chat window to allow the two human participants on the same team to collaborate. Each team member plays in his or her simulation, and the progress of the simulation is synchronized by items among the team members. It takes about 50 min for each session to complete, and the time-stamped chats are the primary source of evidence for the CPS constructs we are targeting. A set of structured system prompts was implemented to facilitate the collaboration (Liu et al., 2015). For each question in the task, the system prompts each team member to respond individually at first. Then, the system prompts the participants to collaborate to discuss their answers. After the collaboration, each member is given a chance to revise his or her initial answer. The difference between the scores on the initial and revised answers captures the gain of the person from the collaboration. The collaboration is considered effective if the sum of the score change is positive, and ineffective otherwise (Hao, Liu, et al., 2016).

We collected the data through a crowdsourcing data collection platform, Amazon Mechanical Turk (Kittur, Chi, & Suh, 2008). We recruited 1,500 participants located in the United States with at least 1 year of college education. We administered to them the general science test, personality survey, and demographic survey. Then we randomly selected 500 participants to complete the single-user version of the simulation. The remaining 1,000 participants 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 communications between the team members around each item. Of the participants, 78% 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 in the simulation tasks were scored based on the scoring rubrics in Zapata-Rivera et al. (2014). The chat communications were annotated into four categories of CPS based on our CPS framework (Liu et al., 2015). In Table 2, we show some example chats and annotations. 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 interrater agreement in terms of unweighted kappa is .67.

Application of the Scoring Strategy to ECSAP Data

In ECSAP, the participants have not been recruited with clear group criteria, such as different states or different companies. However, we can always classify the teams into effective collaboration teams and ineffective collaboration teams

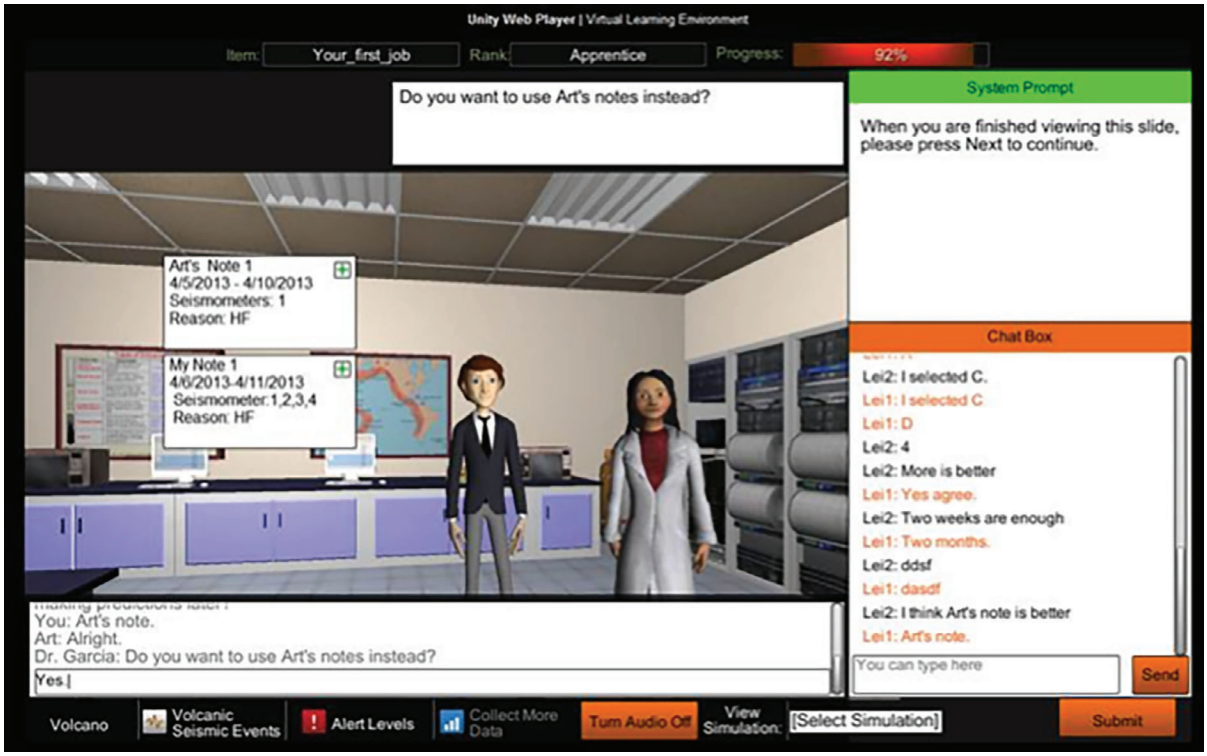

Figure 3 Simulation-based collaborative task used in the ETS Collaborative Science Assessment Prototype.

Table 2 Example of a Part of Annotated Data From One Team

Topic	Chat	Coding	Code meaning
IntroduceYourselves	hello	3	Maintaining
IntroduceYourselves	hey	3	Maintaining
Question1A	chose b, cause its rocks cracking that cause the high frequency events	0	Sharing
Question1A	yes, same here	1	Negotiating
Question1B	d sound right to you?	2	Regulating
Question1B	I couldn't remember, I thought it was C	2	Regulating
Question1B	you are right	1	Negotiating
QuestionsP2	A and B?	2	Regulating
QuestionsP2	yes, that's what i got	1	Negotiating
QuestionsP3	52431?	2	Regulating
QuestionsP3	I was only sure about 5 and 1 being first and last	0	Sharing
QuestionsP3	4 is probably second to last	0	Sharing
ExampleSeisQuestion1	A?	2	Regulating
ExampleSeisQuestion1	picked a	0	Sharing
ExampleSeisQuestion2	thoughts?	2	Regulating
ExampleSeisQuestion2	b?	2	Regulating
ExampleSeisQuestion2	same	1	Negotiating
ExampleSeisQuestion3	obviously c	0	Sharing
ExampleSeisQuestion3	c	0	Sharing

Note. The topic column indicates the specific items around which the conversations happened.

based on their collaboration outcomes (as described in the previous section). An advantage of using the outcome-based grouping criteria is that we automatically get a mapping between the collaboration process and outcomes when we compare the group-wise difference of the process-based features. Hao, Liu, et al. (2016) have explored the mapping between annotation-based features (e.g., Set I features) and the outcomes and found that the negotiation-related skill is crucial for achieving effective collaboration. In this report, we further extend that work by including more extended features from Sets II and III.

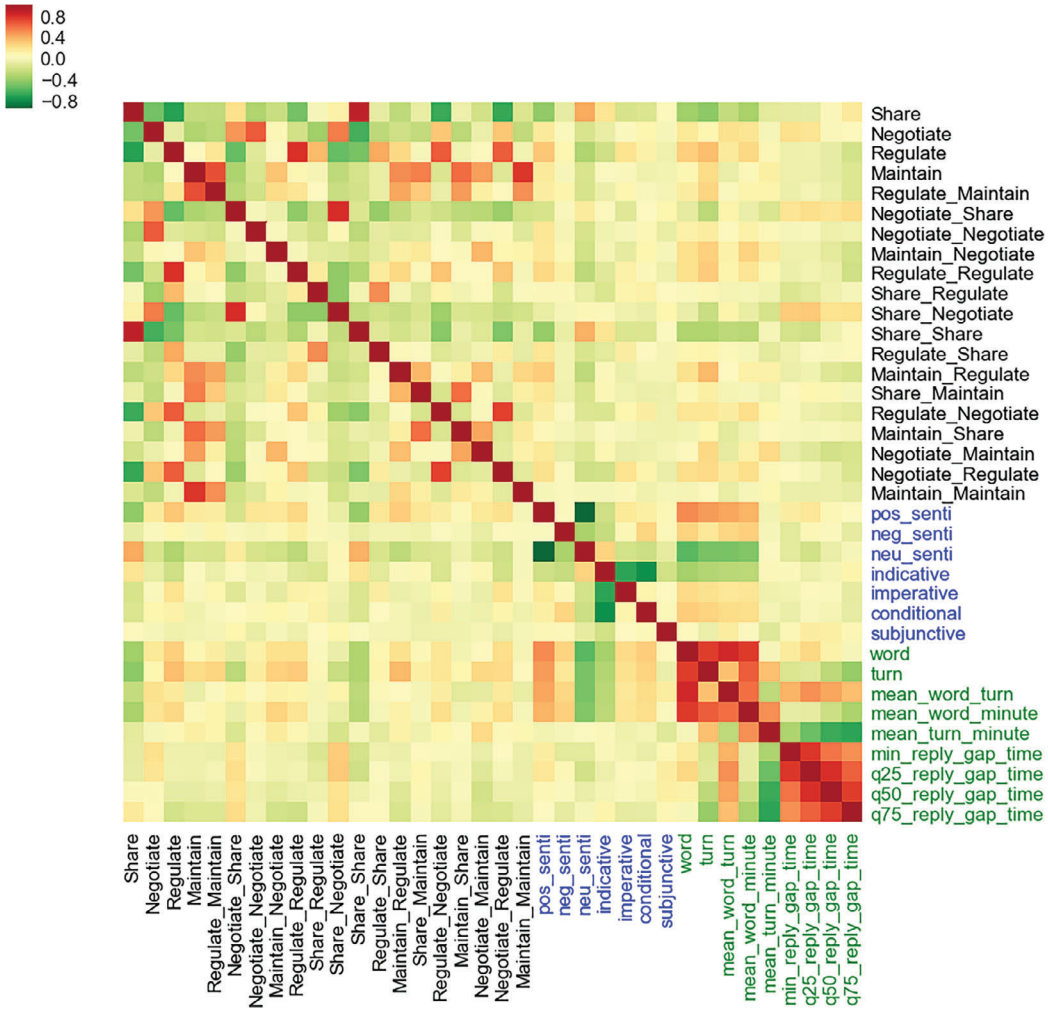

Figure 4 Correlations among the features that represent the collaboration process. The black, blue, and green colors indicate the Set I, II, and III features, respectively.

We aimed to collect CPS data from 500 dyads, as shown in Figure 2. After removing the teams that did not complete the task, we were left with 482 dyads. In each team's response, there are approximately 80 turns of chats, on average, and approximately 15 questions in the simulation-based task. For the first seven selected-response questions, each team member answers the question individually at first, then discusses and collaborates, and then, finally, revises his or her initial answer individually. We consider the collaboration to be effective if the total changes between revised and initial answers for the team are positive.[4] In the following, we focus on the team discussions around the first seven questions, where approximately 30 turns of discussion occurred on average for each team. We noticed that many teams did not precisely follow the initial-collaborate-revise procedure and started some nonprompted discussions when they were asked to answer alone. In our analysis, we consider only the teams that have no more than two nonprompted discussions. After this cut, we are left with 241 dyads out of the 482 dyads. Among the 241 dyads, 154 were classified as effective collaboration, and 87 were classified as ineffective collaboration, based on the score changes between the initial and revised responses. All our analyses will be on this subset.

We first examine how the features (as shown in Table 1) correlate with each other; the results are shown in Figure 4. The colors of the cells indicate the correlation strengths between the features on the rows and columns. Different sets of features are assigned with labels of different colors. Several subsets of the features are highly correlated. For example, the timing-based features (e.g., q25_reply_gap_time) are highly correlated. The word- and turn-related features are also highly correlated. One can expect that a factor analysis may reveal some latent structures from these features, which is not the focus of the current report and will be left for future exploration.

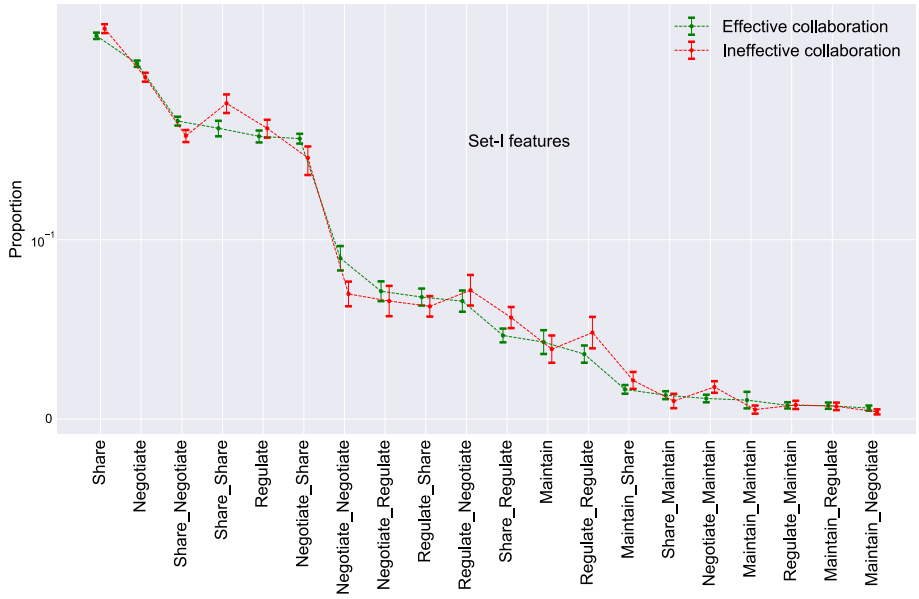

Figure 5 The difference of the mean Set I feature values for the groups with effective (green) and ineffective (red) outcomes. The error bar is the standard error of the mean.

Next, we check how the features are different between the group of teams with effective and ineffective collaboration. A straightforward way is an ANOVA/MANOVA analysis to test the differences. However, these kinds of analyses have low power relative to properly adjusted pairwise *t*-tests and shed little light on which features are important for the outcome measures. So, following the work in Hao, Liu, et al. (2016), we directly examine the feature-wise difference. The results are shown in Figures 5–7. Looking at these plots, one can readily identify that the teams with effective collaboration show more negotiation-related skills, more positive sentiments, more words, and turns. Given that the feature values span a varied scale, the difference is not very salient on the plots. In Figure 8, we show the *p*-values from a pairwise *t*-test of the features from the two groups.[5] A .05 significance level (red dashed line) is placed on the plot to highlight which features are "significantly" different between the two groups. On this basis, the fraction of negotiating, share–share, and share–negotiate; the total number of words; the fraction of positive sentiment; and the fraction of neutral sentiment stand out. By combining the results from Figures 5–7, we further note that the effective-collaboration group shows higher feature values, except for the share–share and neutral sentiment features, which is very consistent with previous findings (Hao, Liu, et al., 2016).

So far, we have examined how different features differ for the groups with different collaboration outcomes. To establish a mapping between the features and outcomes, we need to build a model to relate the process variables and outcome variables. In our case, we have 1 outcome variable that has 2 categories, effective and ineffective, and 36 feature variables corresponding to the collaboration process. In particular, we consider three machine learning models: support vector machine, maximum entropy (also known as logistic regression in statistics literature), and random forest. We evaluate the mapping by the prediction of the outcomes based on threefold cross-validation. The average precision, recall, and F1 scores are shown in Table 3. These results show that one can expect a decent prediction of the collaboration outcomes based on the process features. Note that we are not attempting to build a statistical inference model here but rather to get a sense of how predictive the process-based features is for the collaboration outcome. Though developing a statistical inference model between the process and outcome should be the ultimate goal, we will not pursue it in the current report so as not to deviate from the main intended goals of this report.

Discussion

Although it is widely agreed that CPS is an important 21st-century skill, there is little consensus on how to assess it, particularly at scale and on a regular basis. In this report, we introduced a set of considerations from the psychometric perspective to guide the development of the assessment of CPS and a general scoring strategy that goes beyond characterizing the complex CPS with an oversimplified unidimensional score.

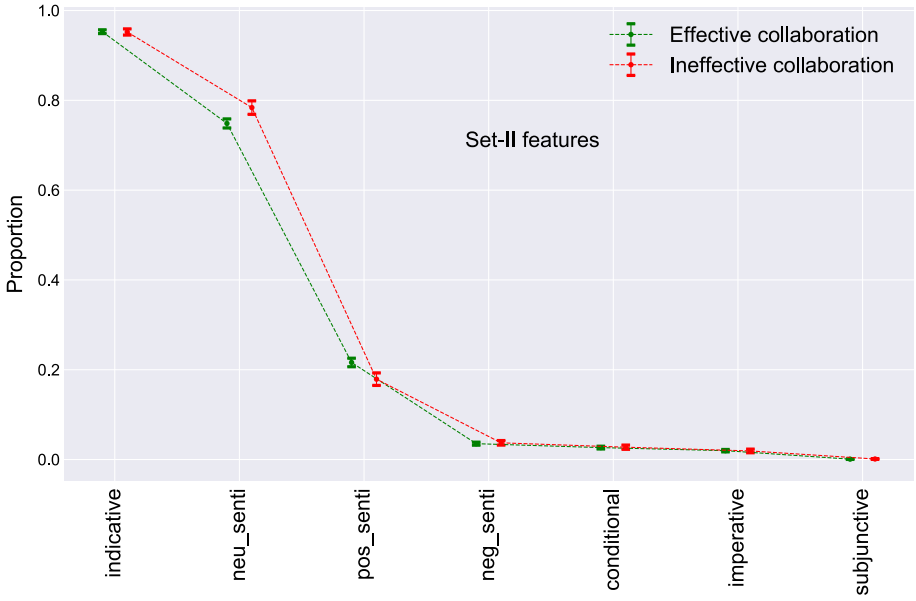

Figure 6 The difference of the mean Set II feature values for the groups with effective (green) and ineffective (red) outcomes. The error bar is the standard error of the mean.

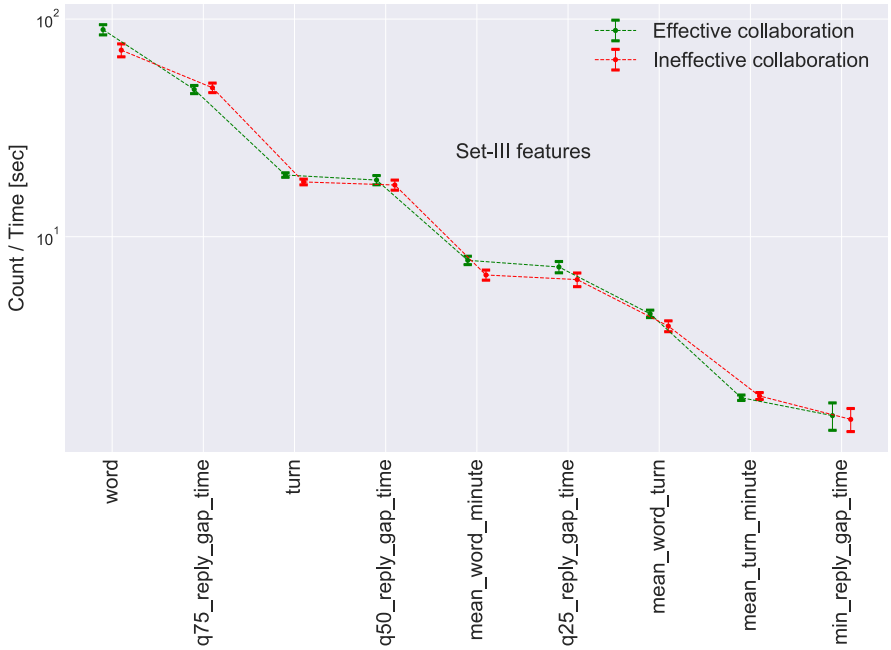

Figure 7 The difference of the mean Set III feature values for the groups with effective (green) and ineffective (red) outcomes. The error bar is the standard error of the mean.

We emphasized that psychometric values like reliability, validity, comparability, and fairness are important considerations in the early design phase (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) and that evidence-centered design is a useful framework for guiding development of a CPS assessment. We noted that two assessment designs were useful for measuring an individual's CPS skill: (a) multiteam round robin-like designs in which an individual participates in different teams and (b) virtual-partner designs that standardize the assessment through the partner's scripts. We pointed out that the reporting goal, either at an individual level or at group level, will determine the choice of assessment design. Group-level reporting may also be useful when only higher level comparisons are needed, for example, at the classroom or company division levels.

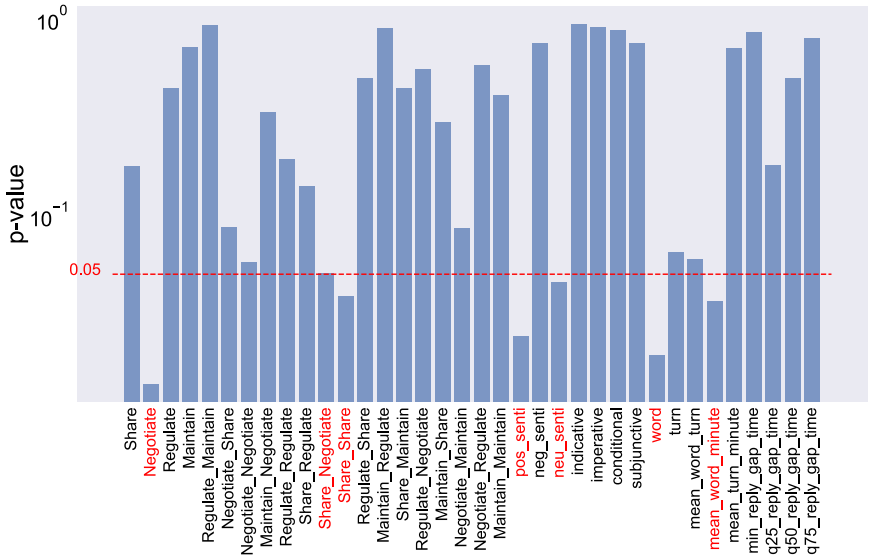

Figure 8 Pairwise *t*-test *p*-values of the features from the group with effective collaboration versus those from ineffective collaboration.

Table 3 Comparison of Performance of Three Methods for Predicting the Collaboration Outcome Based on the Process Features

Method	Precision	Recall	F1
Support vector machine	.64	.95	.77
Logistic regression	.64	.90	.75
Random forest	.66	.73	.70

Note. The results are based on a threefold cross-validation.

We proposed a general scoring strategy for evaluating CPS that considers both outcome and process measures and the statistical/machine learning techniques appropriate for examining the process–outcome relationship. Process measures are based on semantic, lexical, and paralinguistic analyses of communications between collaborators during problem solving and could include features like word complexity, sentiment, grammatical mood, turn taking, and communication volume. We illustrated these ideas with ECSAP administered to 500 individuals and 500 dyads. Comparing processes used by teams with effective versus ineffective collaboration outcomes, we found several differentiators, such as effective teams engaged in more communication and negotiation, and exchanges were more likely positive or neutral compared to ineffective teams. We also showed that process features were able to predict collaborative outcomes.

We believe that the methods and findings discussed here are useful to the broader collaborative learning and collaborative assessment communities. The specific approaches for annotating chats, categorizing them to a CPS taxonomy, and employing statistical methods to evaluate their connection to collaborative problem-solving outcomes are useful. However, we also believe that sustained effort around some of the broader themes addressed here, such as the importance of sound design and analysis approaches and the use of general features and skills taxonomies that become increasingly standardized with experiences in various collaborative domains, are equally important.

A practical challenge for developing assessments of CPS is that it often takes multiple years to finish a cycle from the task design and development to data collection, analysis, and reporting of the findings. As such, we have structured our efforts into several stages. The empirical study and the corresponding findings, as reported here, are from Stage I of our efforts toward developing serious assessments for CPS. As such, they bear some known limitations. For example, we restricted ourselves to a subset of CPS constructs in online collaborations rather than face-to-face collaborations; we considered only the CPS constructs in a specific domain, science; and we chose only text-mediated communications due to both privacy concerns and budget constraints. We targeted only the group-level reporting to simplify the assessment design to make it feasible under the budget and time constraints. In our ongoing Stage II efforts along this line of research, we have developed an online platform, EPCAL (Hao, Liu, von Davier, Lederer, et al., 2017), to facilitate the large-scale study of CPS. EPCAL supports video communication in addition to text chats and has the flexibility to wrap in different

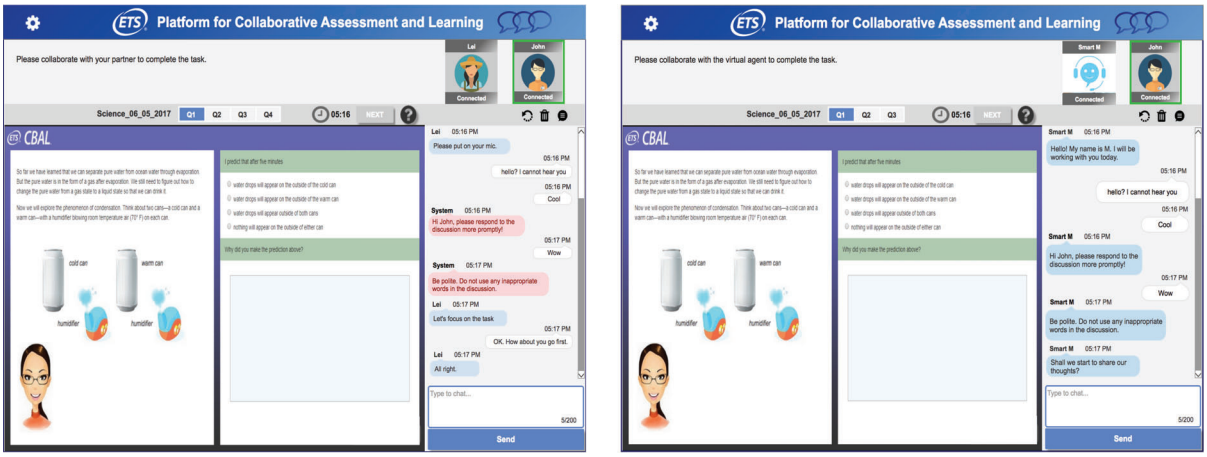

Figure 9 (Left) Human–human and (right) human–agent collaboration supported by the ETS Platform for Collaborative Assessment and Learning.

tasks easily. The easy interchangeability of tasks allows us to explore CPS across a wide range of domains and makes it possible to study the reliability of the assessment by parallel tasks. Moreover, EPCAL supports both human–human and human–agent collaboration (e.g., see Figure 9 for screenshots of the two conditions), which gives us more flexibility to implement different assessment designs for both group-level and individual-level reporting purposes. We will report the relevant work and findings in the near future.

ACKNOWLEDGMENT

Alina von Davier contributed to this report while on staff at Educational Testing Service. She is now with ACTNext. This work was funded in part by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) under Grant # W911NF-19-1-0106 and in part by the ETS Research Allocation. The views, opinions, and conclusions contained in this document are those of the authors and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documents.

Notes

1 Note that proper procedures need to be carried out to correct for possible learning effects if the assessment is done through a long-term span.
2 In fact, simple random sampling is not enough. Instead, a proper sampling design needs to be implemented to reflect the population differences.
3 We use these three sets of features as examples in this report but would like readers to keep in mind that there could be other features that are not covered by these three sets.
4 In the case that the initial answers are correct, there will be no room for improvement. This means that the corresponding item or question is not a good one for eliciting evidence of CPS of the particular team and its members. Just as in most standardized testing programs, selecting appropriate items is an essential step in assessment assembly. A similar situation holds for the assessment of CPS.
5 Note that our purpose here is to identify the features that show significant difference between the two groups rather than statistically concluding how different the two groups are, so we ignore the Bonferroni correction that may be needed for the latter purpose.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrews, J. J., Kerr, D., Mislevy, R. J., Davier, A., Hao, J., & Liu, L. (2017). Modeling collaborative interaction patterns in a simulation-based task. *Journal of Educational Measurement*, *54*, 54–69. https://doi.org/10.1111/jedm.12132

Andrews-Todd, J., Jackson, G. T., & Kurzum, C. (2019). *Collaborative problem solving assessment in an online mathematics task* (Research Report No. RR-19-24). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12260

Andrews-Todd, J., & Kerr, D. (2019). Application of ontologies for assessing collaborative problem solving skills. *International Journal of Testing*, *19*, 172–187. https://doi.org/10.1080/15305058.2019.1573823

Dingler, C., von Davier, A. A., & Hao, J. (2017). Methodological challenges in measuring collaborative problem-solving skills over time. In E. Salas, W. B. Vessey, & L. B. Landon (Eds.), *Team dynamics over time* (pp. 51–70). Bradford, England: Emerald.

Dönmez, P., Rosé, C., Stegmann, K., Weinberger, A., & Fischer, F. (2005). Supporting CSCL with automatic corpus analysis technology. In T. Koschmann, D. D. Suthers, & T.-W. Chan (Eds.), *Computer supported collaborative learning 2005: The next 10 years! Proceedings of the International Conference* (pp. 125–134). Mahwah, NJ: Lawrence Erlbaum.

Dowell, N. M., Graesser, A. C., & Cai, Z. (2016). Language and discourse analysis with Coh-Metrix: Applications from educational material to learning environments at scale. *Journal of Learning Analytics*, *3*(3), 72–95. https://doi.org/10.18608/jla.2016.33.5

Fiore, S. M., Graesser, A., Greiff, S., Griffin, P., Gong, B., Kyllonen, P., … von Davier, A. (2017). *Collaborative problem solving: Considerations for the National Assessment of Educational Progress*. Washington, DC: National Center for Education Statistics.

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 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 31–41). Stroudsburg, PA: Association for Computational Linguistics.

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*, 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Graesser, A. C., Cai, Z., Hu, X., Foltz, P. W., Greiff, S., Kuo, B.-C., … Shaffer, D. (2017). Assessment of collaborative problem solving. In R. Sottilare, A. Graesser, X. Hu, & G. Goodwin (Eds.), *Design recommendations for intelligent tutoring systems: Volume 5. Assessment* (pp. 275–285). Orlando, FL: U.S. Army Research Laboratory.

Graesser, A. C., Foltz, P. W., Rosen, Y., Shaffer, D. W., Forsyth, C., & Germany, M.-L. (2018). Challenges of assessing collaborative problem solving. In E. Care, P. Griffin, & M. Wilson (Eds.), *Assessment and teaching of 21st century skills* (pp. 75–91). New York, NY: Springer.

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. https://doi.org/10.1086/678293

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

Häkkinen, P., Järvelä, S., & Mäkitalo, K. (2003). Sharing perspectives in virtual interaction: Review of methods of analysis. In B. Wasson, S. Ludvigsen, & U. Hoppe (Eds.), *Designing for change in networked learning environments* (pp. 395–404). New York, NY: Springer.

Halpin, P. F., von Davier, A. A., Hao, J., & Liu, L. (2017). Measuring student engagement during collaboration. *Journal of Educational Measurement*, *54*, 70–84. https://doi.org/10.1111/jedm.12133

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* (Research Report No. RR-17-58). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12184

Hao, J., Liu, L., von Davier, A., & Kyllonen, P. (2015). Assessing collaborative problem solving with simulation based tasks. In O. Lindwall, P. Hakkinen, T. Koschmann, P. Tchounikine, & S. Ludvigsen (Eds.), *Exploring the material conditions of learning: The computer supported collaborative learning (CSCL) conference* (2nd ed., pp. 544–547). Philadelphia, PA: International Society of the Learning Sciences.

Hao, J., Liu, L., von Davier, A., & Kyllonen, P. C. (2017). Initial steps towards a standardized assessment for collaborative problem solving (CPS): Practical challenges and strategies. In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 135–156). New York, NY: 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. In T. Barnes, M. Chi, & M. Feng (Eds.), *EDM16: Proceedings of the 9th International Conference on Educational Data Mining* (pp. 382–387). Worcester, MA: International Educational Data Mining Society.

Hao, J., Liu, L., von Davier, A. A., Lederer, N., Zapata-Rivera, D., Jakl, P., & Bakkenson, M. (2017). *EPCAL: ETS platform for collaborative assessment and learning* (Research Report No. RR-17-49). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12181

Hao, J., & Mislevy, R. J. (2019). Characterizing interactive communications in computer-supported collaborative problem-solving tasks: A conditional transition profile approach. *Frontiers in Psychology*, *10*, 1–9. https://doi.org/10.3389/fpsyg.2019.01011

Hao, J., Smith, L., Mislevy, R., von Davier, A., & Bauer, M. (2016). *Taming log files from game/simulation-based assessments: Data models and data analysis tools* (Research Report No. RR-16-10). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12096

He, Q., von Davier, M., Greiff, S., Steinhauer, E. W., & Borysewicz, P. B. (2017). Collaborative problem solving measures in the Programme for International Student Assessment (PISA). In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 95–111). New York, NY: Springer.

Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (in press). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2018.07.035

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and research* (pp. 37 – 56). New York, NY: Springer.

Jeong, H., & Hmelo-Silver, C. E. (2010). Technology use in CSCL: A content meta-analysis. In *2010 43rd Hawaii International Conference on System Sciences (HICSS)* (pp. 1 – 10). Piscataway, NY: IEEE.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In M. Burnett, M. F. Costabile, T. Catarci, B. de Ruyter, D. Tan, M. Czerwindsky, & A. Lund (Eds.), *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 453 – 456). New York, NY: ACM.

Liu, L., Hao, J., Andrews, J. J., Zhu, M., Mislevy, R. J., Kyllonen, P., … Graesser, A. (2017). *Collaborative problem solving: Innovating standardized assessment*. Philadelphia, PA: International Society of the Learning Sciences.

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

Martin-Raugh, M. P., Kyllonen, P. C., Hao, J., Bacall, A., Becker, D., Kurzum, C., … Barnwell, P. (in press). Negotiation as an interpersonal skill: Generalizability of negotiation outcomes and tactics across contexts at the individual and collective levels. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2019.03.030

Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13 – 23. https://doi.org/10.1002/j.2333-8504.1992.tb01470.x

Mislevy, R. J., & Riconscente, M. (2006). Evidence-centered assessment design. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 61 – 90). Mahwah, NJ: Erlbaum.

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

Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In C. O'Malley (Ed.), *Computer supported collaborative learning* (pp. 69 – 97). New York, NY: Springer.

Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, *3*, 237 – 271. https://doi.org/10.1007/s11412-007-9034-0

Rosen, Y. (2014). Comparability of conflict opportunities in human-to-human and human-to-agent online collaborative problem solving. *Technology, Knowledge and Learning*, *19*, 147 – 164. https://doi.org/10.1007/s10758-014-9229-1

Rosen, Y. (2015). Computer-based assessment of collaborative problem solving: Exploring the feasibility of human-to-agent approach. *International Journal of Artificial Intelligence in Education*, *25*, 380 – 406. https://doi.org/10.1007/s40593-015-0042-3

Rosen, Y. (2018). Computer agent technologies in collaborative learning and assessment. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (4th ed., pp. 2402 – 2410). Hershey, PA: IGI Global.

Rundgren, C.-J., Rundgren, S.-N. C., Tseng, Y.-H., Lin, P.-L., & Chang, C.-Y. (2012). Are you slim? Developing an instrument for civic scientific literacy measurement (SLIM) based on media coverage. *Public Understanding of Science*, *21*, 759 – 773. https://doi.org/10.1177/0963662510377562

Rus, V., Moldovan, C., Niraula, N., & Graesser, A. C. (2012, June). *Automated discovery of speech act categories in educational games*. Paper presented at the International Conference on Educational Data Mining, Chania, Greece.

Scoular, C., Care, E., & Awwal, N. (2017). An approach to scoring collaboration in online game environments. *Electronic Journal of e-Learning*, *15*(4). Retrieved from http://www.ejel.org/volume15/issue4/p335

Stahl, G., Koschmann, T., & Suthers, D. (2006). Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 409 – 426). Cambridge, England: Cambridge University Press.

Strijbos, J.-W., Martens, R. L., Prins, F. J., & Jochems, W. M. (2006). Content analysis: What are they talking about? *Computers & Education*, *46*, 29 – 48. https://doi.org/10.1016/j.compedu.2005.04.002

Suthers, D. D., Hundhausen, C. D., & Girardeau, L. E. (2003). Comparing the roles of representations in face-to-face and online computer supported collaborative learning. *Computers & Education*, *41*, 335 – 351. https://doi.org/10.1016/j.compedu.2003.04.001

Tutty, J. I., & Klein, J. D. (2008). Computer-mediated instruction: A comparison of online and face-to-face collaboration. *Educational Technology Research and Development*, *56*, 101 – 124. https://doi.org/10.1007/s11423-007-9050-9

von Davier, A. A., & Halpin, P. F. (2013). *Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations* (Research Report No. RR-13-41). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02348.x

von Davier, A. A., Hao, J., Liu, L., & Kyllonen, P. (2017). Interdisciplinary research agenda in support of assessment of collaborative problem solving: Lessons learned from developing a collaborative science assessment prototype. *Computers in Human Behavior*, *76*, 631 – 640. https://doi.org/10.1016/j.chb.2017.04.059

von Davier, A. A., Zhu, M., & Kyllonen, P. C. (2017). *Innovative assessment of collaboration*. New York, NY: Springer.

Warkentin, M. E., Sayeed, L., & Hightower, R. (1997). Virtual teams versus face-to-face teams: An exploratory study of a Web-based conference system. *Decision Sciences*, *28*, 975–996. https://doi.org/10.1111/j.1540-5915.1997.tb01338.x

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, *46*, 71–95. https://doi.org/10.1016/j.compedu.2005.04.003

World Economic Forum. (2015). *New vision for education: Unlocking the potential of technology*. Vancouver, BC: British Columbia Teachers' Federation.

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. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Intelligent tutoring systems: Proceedings of the 12th International Conference, ITS 2014* (pp. 625–626). Cham, Switzerland: Springer.

Zhu, M., & Andrews-Todd, J. (2019, June). *Understanding the connections of collaborative problem solving skills in a simulation-based task through network analysis*. Paper presented at the International Conference on Computer Supported Collaborative Learning, Lyons, France.

Zhu, M., & Zhang, M. (2017). *Network analysis of conversation data for engineering professional skills assessment* (Research Report No. RR-17-59). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12191

Suggested citation:

Hao, J., Liu, L., Kyllonen, P., Flor, M., & von Davier, A. A. (2019). *Psychometric considerations and a general scoring strategy for assessments of collaborative problem solving* (Research Report No. RR-19-41). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12276

Action Editor: Gautam Puhan

Reviewers: Andreas Oranje and Mengxiao Zhu

REPORT DOCUMENTATION PAGE

1. REPORT DATE (Month Year)	2. REPORT TYPE	3. DATES COVERED (Month Year)	
December 2019	Technical Report	START DATE January 2018	END DATE December 2018

4. TITLE AND SUBTITLE

Psychometric Considerations and a General Scoring Strategy for Assessments of Collaborative Problem Solving

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. COOPERATIVE AGREEMENT NUMBER
	W911NF-19-1-0106	

5d. PROGRAM ELEMENT NUMBER	5e. PROJECT NUMBER	5f. TASK NUMBER	5g. WORK UNIT NUMBER

6. AUTHOR(S)

Hao, Jiangang, Liu, Lei, Kyllonen, Patrick, Flor, Michael, & von Davier, Alina A.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541

8. PERFORMING ORGANIZATION REPORT NUMBER

ETS RR–19-41

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
U.S. Army Research Institute for the Behavioral and Social Sciences 6000 6th Street (Bldg. 1464 / Mail Stop: 5610) Fort Belvoir, Virginia 22060-5610	ARI	

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution Statement A. Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES This Research Report is part of the ETS Research Report Series (ISSN 2330-8516)
Online 06 November 2019 DOI: https://doi.org/10.1002/ets2.12276

14. ABSTRACT

Collaborative problem solving (CPS) is an important 21st-century skill that is crucial for both career and academic success. However, developing a large-scale and standardized assessment of CPS that can be administered on a regular basis is very challenging. In this report, we introduce a set of psychometric considerations and a general scoring strategy around assessing CPS, summarized based on the results of the extensive empirical studies we conducted at Educational Testing Service (ETS) over the past 6 years. Using the ETS Collaborative Science Assessment Prototype as an example, we show how these psychometric considerations have been incorporated into the development of the assessment prototype and how the scoring strategy has been implemented.

15. SUBJECT TERMS

Collaborative problem solving; psychometrics; assessment design; scoring

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
Unclassified	Unclassified	Unclassified	Unlimited Unclassified	17

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER (Include area code)
Dorothy Young	703-545-2316