International Journal of Assessment Tools in Education
2023, Vol. 10, Special Issue, 194–207
https://doi.org/10.21449/ijate.1407315
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Collaborative problem-solving design in large-scale assessments: Shedding lights in sequential conversation-based measurement

Qiwei He [1*]

[1]Georgetown University, Data Science and Analytics Program, Washington, DC, 20057 USA

ARTICLE HISTORY

Received: Dec. 20, 2023
Accepted: Dec. 24, 2023

Keywords:
Collaborative-problem solving,
Conservation path,
Sequential measurement
PISA,
Item design.

Abstract: Collaborative problem solving (CPS) is inherently an interactive, conjoint, dual-strand process that considers how a student reasons about a problem as well as how s/he interacts with others to regulate social processes and exchange information (OECD, 2013). Measuring CPS skills presents a challenge for obtaining consistent, accurate, and reliable scale across individuals and user populations. The Programme for International Student Assessment (PISA)'s 2015 cycle first introduced an assessment of CPS in international large-scale assessments in which computer-based conversational agents were adapted to represent team members with a range of skills and abilities. This study draws on measures of the CPS domain in PISA 2015 to address the challenges and solutions related to CPS item design and shed lights on sequential conversation-based measurement. Specifically, we present the process of CPS item design, the development of scoring rules through CPS conversation paths, and discuss the possible approaches to better estimate CPS beyond item response models.

1. LANGUAGE MODELS IN AUTOMATED ESSAY SCORING

Researchers consider the importance of collaborative problem solving as an educational outcome and a skill for life and work as having increased since the turn of the 21st century (National Center for Educational Statistics, 2015; National Academies, 2012; Wildman et al., 2012; Casner-Lotto & Barrington, 2006). Noncognitive skills that intersect with cognitive ones now involve mastering new challenges and require cooperative efforts among a group of individuals (Griffin et al., 2012; Greiff et al., 2014). Collaborative problem solving (CPS) is inherently an interactive, conjoint, dual-strand process that considers how the student reasons about a problem as well as how he or she interacts with others to regulate social processes and exchange information (Organisation for Economic Co-operation and Development [OECD], 2013). While measuring CPS skills presents a challenge for obtaining consistent, accurate, and reliable measurement across individuals and across user populations, it is an also opportunity to gain more information about cognitive processes in interactions with peers. (He et al., 2017).

As Stecher and Hamilton (2004) observed, CPS skills are difficult to measure. Challenges persist from two major aspects: first, developing items with complex constrains, and second, producing a reliable scale to measure the CPS skills in an accurate way. Given concerns about

*CONTACT: Qiwei He qiwei.he@georgetown.edu Georgetown University, Faculty of Education, Department of Educational Sciences, Washington, DC, 20057 USA

e-ISSN: 2148-7456 /© **IJATE** 2023

language factors and fairness across different countries and cultures, even more difficulties have to be confronted when measuring CPS skills in large-scale assessments such as the Programme for International Student Assessment (PISA). Traditional methods that have been generally used for item response modeling may not be appropriate for measuring collaborative interactions because of the dependence within elements of complex tasks and between interacting participants (Cooke et al., 2012; Quellmalz et al., 2009). Therefore, new assessment designs and statistical methods that capture the dynamic of knowledge sharing in collaborative contexts are required (Dede, 2012). How to model such knowledge and skills in a way that meets the technical standards of traditional assessments is an issue that urgently needs to be solved.

A bold move was made in PISA 2015 to introduce CPS to the assessment program (OECD, 2013). This objective was accomplished through the successful implementation of conversational agents incorporated in computer-based testing. Such innovation introduced a new perspective to understanding students' performance that goes beyond the borders of domain-specific competencies and mere cognitive ability constructs such as reasoning and working memory (Greiff et al., 2014). Skillfully dealing with new problems in diverse settings and contexts, as part of a team instead of individually, is at the core of the concept of CPS. CPS reflects a set of skills that combines cognitive and social aspects that are relevant for successful problem solving across domains regardless of the specific contextual setting (He et al., 2017).

The triennial PISA study aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-olds. In 2015 over a half million students, representing 28 million across 72 countries and economies, took PISA in three core cognitive domains—science, reading, and math, as well as CPS and financial literacy. PISA has a history of measuring problem-solving skills, specifically individual problem solving in PISA 2003 (paper and pencil based) and 2012 (computer based), acknowledging these skills' increasing relevance.

This study draws on measures of the CPS domain in PISA 2015 to address the challenges and solutions related to CPS item design and shed lights on sequential conversation-based measurement. Specifically, we present the process of CPS item design, the development of scoring rules through CPS conversation paths, and discuss the possible approaches to better estimate CPS beyond item response models.

In the following section, we introduce the process of CPS item design for PISA 2015 and examine factors that potentially make impact on CPS item difficulty. The CPS scoring rules are illustrated through conversation paths in Section 3. We finalized this paper with a general conclusion on reliability of CPS scale in PISA 2015 and some discussions on future research directions for CPS assessments.

2. DEVELOPING CPS ITEMS FOR PISA 2015

2.1. CPS Item Design

For PISA 2015, CPS is defined as follows: "CPS competency is the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution" (OECD, 2013). As such, this competence integrates two essential concepts: problem solving and collaboration, which were categorized into a set of 12 CPS skills. As shown in Table 1, a matrix of CPS skills was created that included three major CPS competencies crossed with four problem-solving processes.

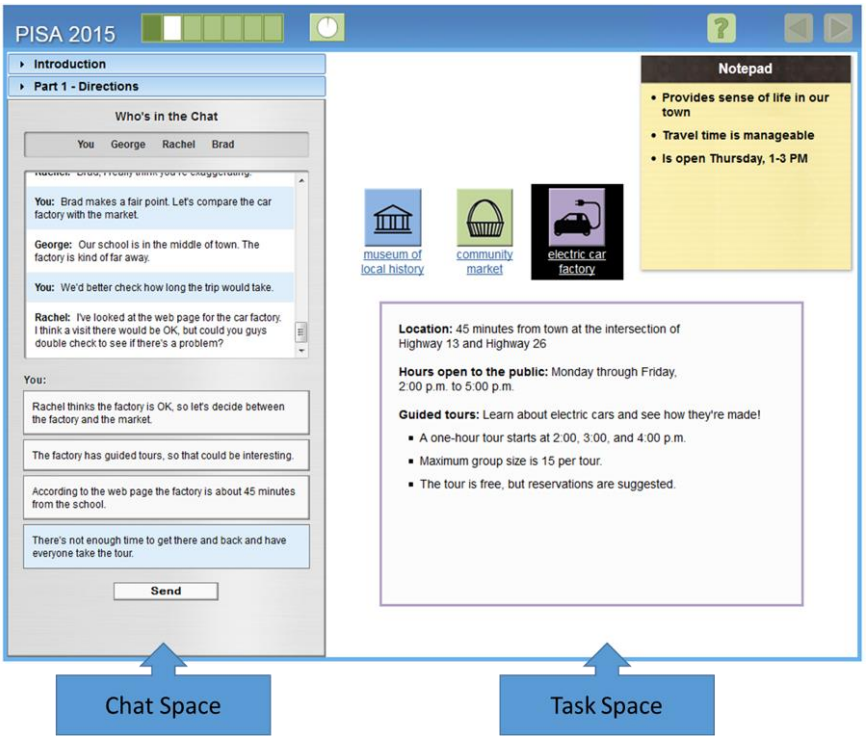
The computer-based CPS tasks (see Figure 1) that were developed to measure these skills were situated in a chat environment ("chat space") where students interacted with one or more simulated agents, identified as teammates, to solve a presented problem. The student was provided with a set of four chat options, and agent responses were based on the option selected.

Each task also included a problem space ("task space") where the student and/or agents could take actions as they worked toward a problem solution. Examples of these actions included selecting information to complete a form or scheduling tasks on a calendar presented in the problem space.

Table 1. *Matrix of Collaborative Problem Solving Skills for PISA 2015 (Organisation for Economic Co-operation and Development, 2013).*

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

Figure 1. *A sample screen of chat and task spaces in a released CPS item (The Visit) in PISA (Organisation for Economic Co-operation and Development, 2015a).*



2.2. Mapping Items onto CPS Skills

As part of the item development process, each item was classified into one of the 12 CPS categories, reflecting the 12 intersecting skills being assessed. Data from the PISA CPS assessment was analyzed to estimate a set of item characteristics for the 117 items included in the main survey.[†] Following data analysis, the items were associated with their difficulty estimates and framework classifications to create an item map. As shown in Table 2, the item map includes information on a certain item along with a brief qualitative description for a subset of CPS items by rows. Table 2 presents two selected items from a released PISA CPS unit (*Xandar*) to illustrate the mapping process, in which the more difficult item is listed first.

Table 2. *A Map for Selected Collaborative Problem-Solving Items in the Released Unit (Xandar).*

Item (Unit and Item ID)	Item Difficulty on CPS Scale	Task Requirements	Establishing & Maintaining Shared Understanding	Taking Appropriate Action to Solve the Problem	Establishing & Maintaining Team Organisation	Exploring & Understanding	Representing & Formulating	Planning and Executing	Monitoring & Reflecting
Xandar CC100203	537	TAKE INITIATIVE by identifying one remaining task needed to solve the problem. Recognize time limits presented in the scenario and assume responsibility for completing the task without further discussion.			○		○		
Xandar CC100301	357	ACT based on agreed-upon role to complete simple assigned task in an uncomplicated problem space.			○			○	

2.3. Examining Factors That Impact CPS Difficulty

The analysis performed to create an item map made it possible to look for factors associated with item difficulty. This could be done by examining the ways in which CPS skills are associated with items located at different points, ranging from the bottom to the top of the scale. When developing a CPS unit, complex constrains may set on items' difficulty level, in order to make a proper mix for items with different difficulty levels. We listed a set of major attributes as below.

2.3.1. *Features of problem complexity*

Features of problem complexity take a high priority in developing a CPS item in accordance with proper difficulty level. There are three major features to help define problem complexity: the nature of the presented problem, the progression of the problem solution, and characteristics of the task space where the problem is worked on.

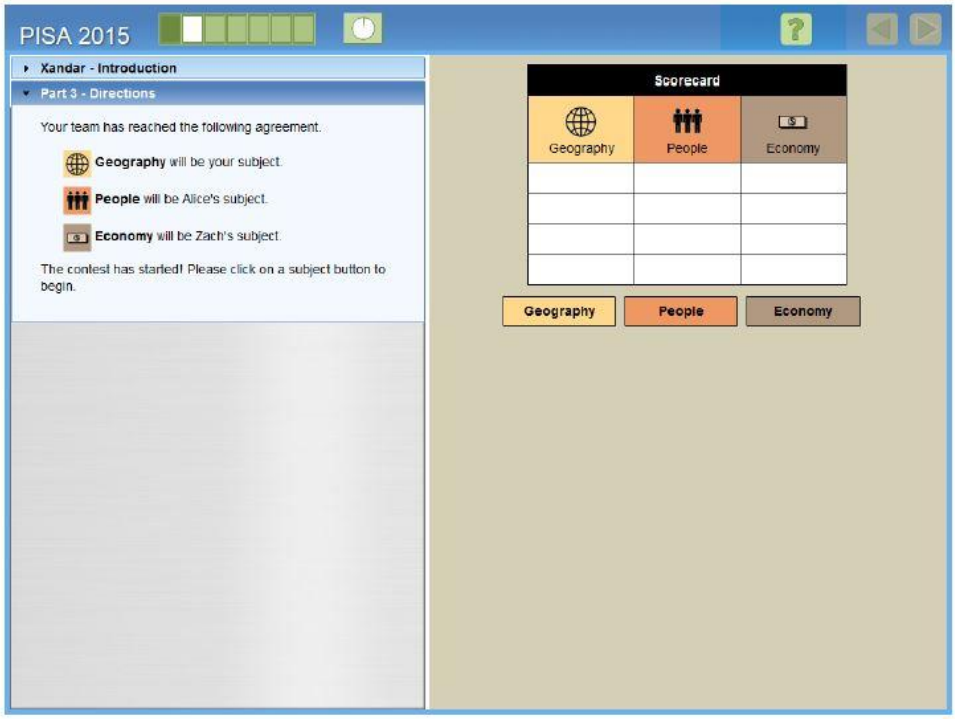
The nature of the presented problem is the first essential element to influence CPS problem complexity. At the lower levels of the scale, problems are well defined with clear goals. Students may be asked to execute a simple and agreed-upon solution, while at higher levels,

[†] This is the number of independently scored items in the final CPS database. Four items included in the main survey were dropped during data analysis. Additionally, a number of items in each unit were combined, based on the main survey analysis and/or to reflect the branching logic within units. As a result of the branching, based on the path students took, students might not see all items in a unit and, therefore, items needed to be clustered in order to function psychometrically.

problems are more complex, requiring students to satisfy multiple constraints, hold more information in working memory, or deal with an impasse or unexpected action. Figure 2 exhibits the screenshot of the lower difficulty item (CC100301) in Table 2. To solve this CPS task, students needed to simply act based on the agreed-upon role, respond to the directions on the screen, and click the correct button. Conversely, the higher difficulty item (CC100203) listed in Table 2, as shown in Figure 3, displayed complexity in the item layout, with an interactive map as well as two tables with dynamic results through the CPS process providing supplementary information. This item required that students respond to a question from one team member and also provided information about how the team is progressing. The additional requirement to identify gaps that had not yet been filled in provides further evidence of its high difficulty level. Students had to use the information displayed in the task space, along with an understanding of how the game worked, to respond correctly.

Figure 2. *A sample CPS item with lower difficulty in a released CPS unit (Xandar) in PISA 2015 (OECD, 2015a).*

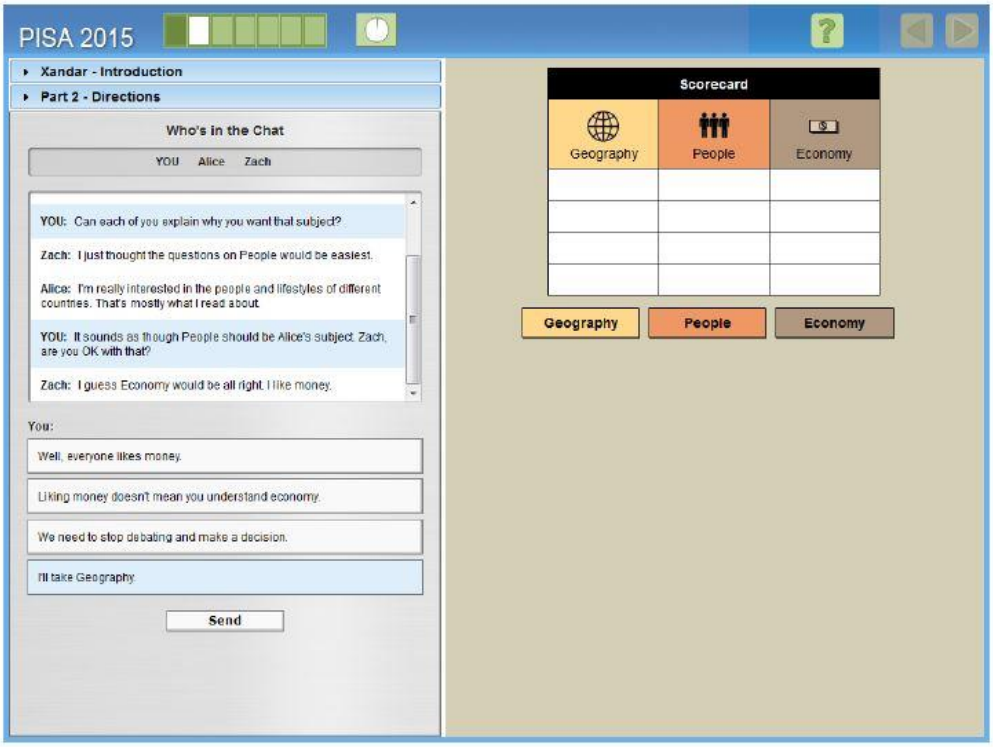
Item	CC100301
Collaborative competency	Establishing and maintaining team organisation
Problem-solving process	Planning and executing
Collaborative problem-solving skill	Following rules of engagement (e.g., prompting other team members to perform their tasks)
Difficulty	357 (Level 1)
Credited action	Student clicks on the "Geography" button.



The second feature of problem complexity relates to the progression of the problem solution. The CPS tasks were chat-based scenarios where information unfolded throughout the course of the task. Item difficulty could therefore be impacted by how recently required information was presented. Having to recall or go back and review information presented earlier in the task makes an item harder to answer. A sequence effect also impacts difficulty in these tasks. Items tend to be easier if they are part of a series of items focusing on a single aspect of the problem and requiring similar student responses.

Figure 3. *A sample CPS item with higher difficulty in a released CPS unit (Xandar) in PISA 2015 (OECD, 2015a).*

Item	CC100203
Collaborative competency	Establishing and maintaining team organisation
Problem-solving process	Representing and formulating
Collaborative problem-solving skill	Describe roles and team organisation (communication protocol/rules of engagement)
Difficulty	537 (Level 2)
Credited response	"I'll take Geography."



Characteristics of the task space are considered the third major feature of problem complexity that influence CPS item difficulty. At lower levels of the scale, changes in the problem space are controlled by the student. Problems may require information to be reordered or new information to be added, but those actions are performed by the respondent. Where information in the problem space changes as a result of agent actions, items tend to be more difficult, particularly in cases where those actions are not explicitly signaled in the chat. In these cases, the student must both notice the changes and infer which of the agents took the action.

Additional aspects of the problem space may affect how difficult it is to solve the presented problem. These include but are not limited to reading load, multiple channels of information-including tables, figures, and diagrams—and the need to use spatial or temporal skills.

2.3.2. *Features of collaboration complexity*

Features of collaboration complexity are often presented by the number of collaborators that are required to be involved in the task and the roles they need to play. In each CPS unit, the student worked with one or more group members to solve a problem, with the group members/computer agents providing input much as fellow students would do. The conversational agents responded to students' textual inputs and actions when the student moved through different stages of the problem. In each stage, communications or actions that could be performed by either the agent or the student was predefined, which resulted in the ability to objectively score all responses.

The computer dynamically monitored the state of the problem through the task completion process. Characteristics of the agents, or virtual team members, with whom the student had to interact also impacted item difficulty. Where agents were collaborative and capable and take an active role in solving the problem, items tended to be easier. In such cases, the student could simply be called upon to provide requested information or agree to the direction being suggested. When agents were focusing on their own goals rather than those of the team, it was more challenging to establish team organization and develop a shared understanding of the problem. The need to collaborate with agents who make errors that need to be noted and remedied can make items more difficult.

The roles of students involved in the collaborative task are also critical to the problem complexity. At the lowest levels of the scale, tasks required only that students respond to agent requests for information or suggestions for actions. More difficult tasks required that students take initiative. That initiative might take several forms including: requesting needed information, suggesting actions for team members to take, and monitoring agent's actions or statements to be sure they are correct and aligned with the agent's agreed-upon role on the team. At higher levels of the scale, tasks required students to resolve a conflict between agents, propose that the team pursue a new approach, or help balance a desire for consensus against the efficiency of the problem-solving process.

2.3.3. *Integration of problem solving and collaboration demands*

Last but not least, the problem solving and collaboration features of each CPS task do not operate in isolation. The difficulty of any given CPS task depends on the interaction between the problem-solving demands and the nature of the collaboration that is required. At the lowest level, items often required either simple collaboration efforts or simple problem-solving tasks. At the highest levels on the scale, complex problem-solving demands were complicated by challenging social interactions, and students had to balance both in order to successfully complete a task.

3. DEFINE CPS SCORING RULE WITH CONVERSATION PATH

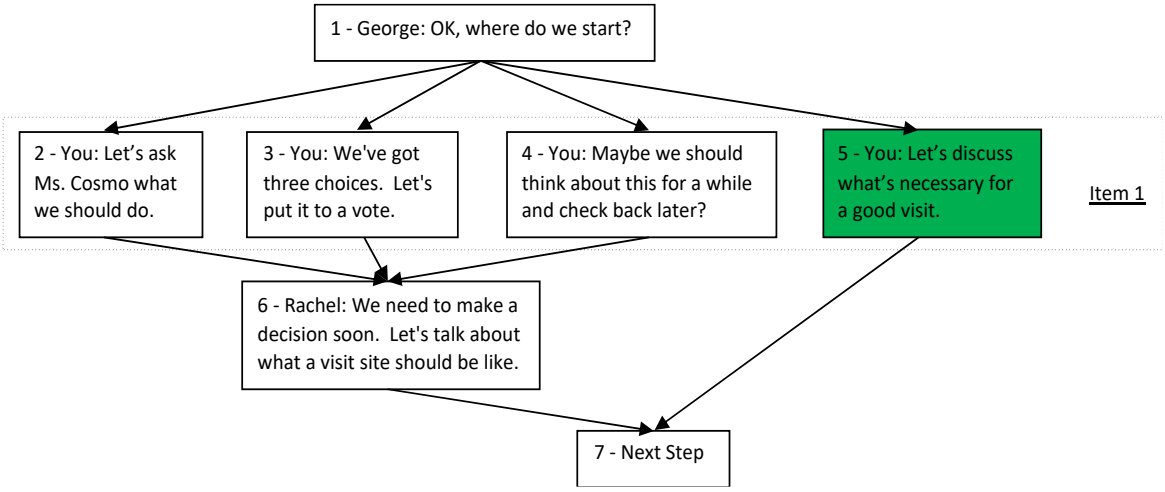
After gaining insight on investigating factors potentially influencing CPS difficulty, we proposed scoring rules for different item types that are specified for the CPS items via path analysis. The construction of different item types was often associated with requirements of different item difficulty levels. To satisfy the conditions of item difficulty level in a specific unit, special item type such as "rescue" and "bonus" items were proposed particularly for the CPS domain. We used some example items here to illustrate how the conversation path analysis worked and how we had to combine items in certain types to keep the CPS scale reliable.

3.1. Conversation Path and Convergence Structure

The major difference between CPS and regular problem solving relates to the perspective of collaboration. In the PISA CPS domain, respondents were required to solve the problem through a collaborative effort, that is, completion of a task with at least two students together rather than an individual alone. As introduced earlier, in one CPS unit, one or many conversational agents worked together with the respondent to go through the dialogues and make "joint" efforts to solve a task. Similar to a computer game, a CPS unit required the respondent to choose an optimal sentence from a set of multiple choices to go through the conversation with agents or choose one or more actions to pass.

Convergence was generally used to guarantee that different paths arrived at an identical point. That is, regardless of what choices the student made, the path led to the same convergence point. Each path to the convergence point had to provide the student with the same information and bring him or her to the same stage of the problem.

Figure 4. *Conversation path of a typical example of CPS item with a simple segment in a released CPS unit (The Visit) in PISA 2015.*



Note. The node highlighted in green is the correct response to this item.

Figure 4 shows an example item with a simple convergence structure in a released CPS unit (The Visit). The collaboration task in this unit was to jointly create a welcoming activity for students coming from abroad. "You" in the script represented the respondent who was required to work with three fellow students (agents)—George, Rachel, and Brad (who shows up later)—to decide what to do to welcome the foreign student. After seeing the input from George (Node 1), the test taker could choose one answer among Node 2 to Node 5 (i.e., Item 1). The respondent got full credit when Node 5 was selected (green). The path then continued to Node 7: the final convergence point. Otherwise, Rachel's response (Node 6) would appear as an intermediate point, and then the path would move on to Node 7 the final convergence point. We defined the phase between two convergence points as one segment, meaning only two convergence points could be found within one segment, the starting convergence point and ending convergence point. A simple segment could have only one scoring item, while a complex segment could have more than one scoring items.

3.2. "Rescue" Items

"Rescue" items were typically developed in a complex segment, where respondents might have the possibility of going through two or three choice points before getting to the convergence point. For example, in The Visit unit, the student and the agents needed to help one of the foreign students get to the airport (see Figure 5 for the item screen and Figure 6 for the conversation path map). The full credited response was the third choice ("I'm at school, where are you guys?"), that is, Node 80 in Figure 6, which told the team his or her location and led directly to the convergence point (Node 85). But students who chose the other paths still arrived at the convergence point, although it took longer. For instance, if the student selected the first option ("What happened to his host family?"), that is, Node 81, Rachel rescued by saying she didn't know what happened to his host family and asking the student if he or she were at school; this gave the student a second chance to choose the response providing his or her location (in Node 83, Node 87, Node 88, and Node 89). If the student selected the second option ("You're good at arranging things, Rachel, can you take care of Zheng?"), that is, Node 78, or the fourth option ("I'm not sure I'm the best person to decide. Rachel, can you help Zheng?"), that is, Node 79, the conversation path worked its way to the final convergence point, meaning students choosing the second and fourth options would not have a chance to answer Item 3. It was noted that the process data in the log file indicated students were unlikely to notice these convergence

and rescue structures. The structure design apparently had little impact on students' test-taking behavior as they progressed through the scenario.

Figure 5. *A sample screen of rescue designs in a released CPS item (The Visit) in PISA (OECD, 2015a).*

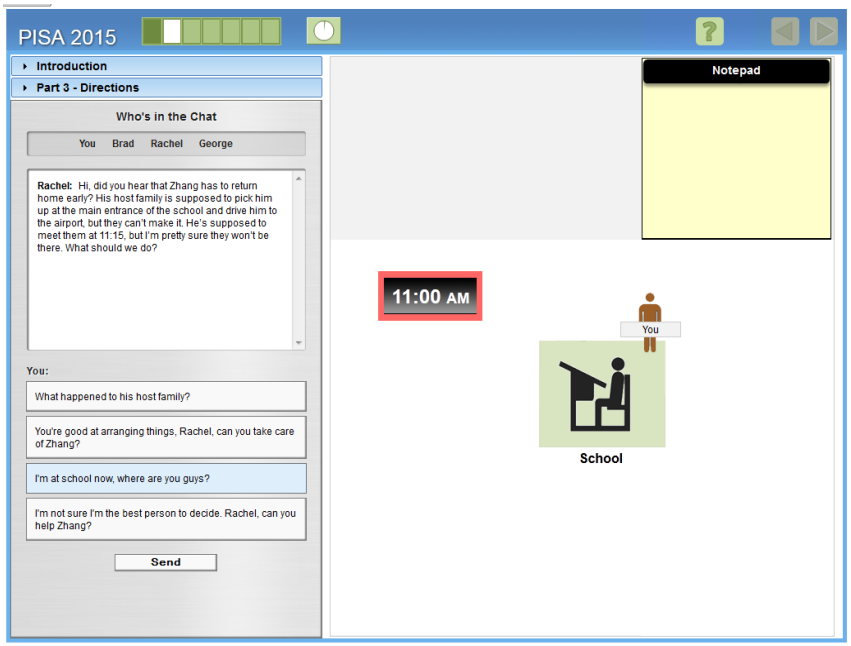
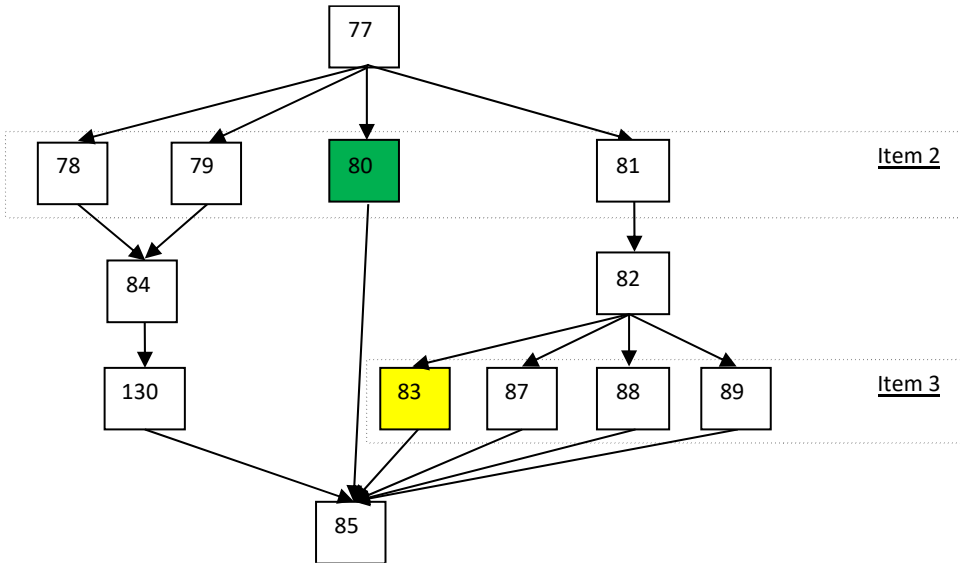


Figure 6. *Conversation path of a "rescue item type" example (see item screenshot in* Figure 5*) of CPS item with a complex segment in a released CPS unit (The Visit) in PISA.*



Note. The node highlighted in green is the correct response to Item 2; the node highlighted in yellow is the correct response to Item 3 on the "rescue" path.

However, the "rescue" item type brought an issue in scoring. Students who got a full credit of 2 points in Item 2 lost the chance to see Item 3, so their Item 3 score was 0; students who got 1 point in Item 3 had already failed in Item 2, recorded as 0 points in Item 2. Therefore, the score correlation between these two items could be substantially negative. One possibility would have been for students who did not have a chance to see Item 3 to receive a score of "not applicable," but such a solution ran counter to the design purpose to assess students' CPS skills based on the selection to the prompt in Node 77.

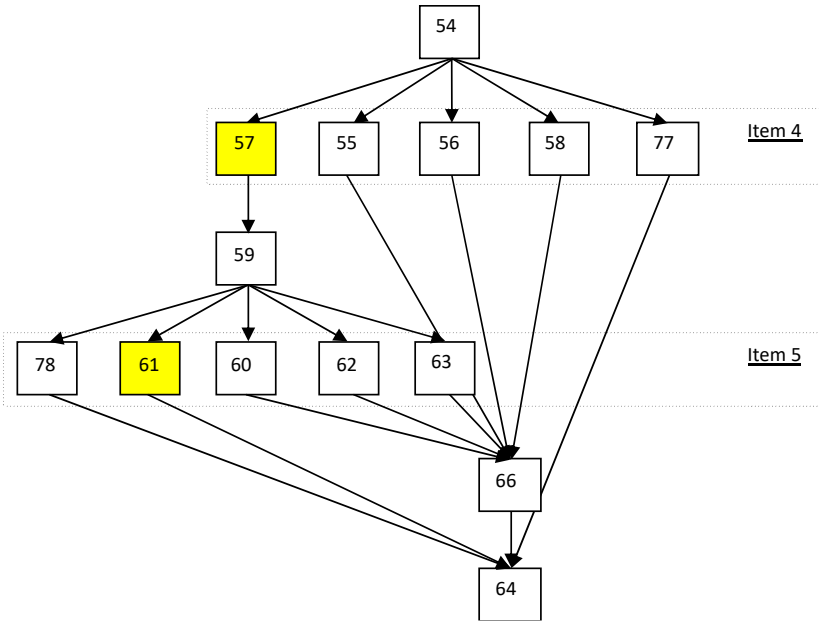
A better solution for such a scoring issue was adopted: to treat the whole complex segment as one polytomous item. Basically, we assigned all item credits within the segment with proper weights. Instead of looking at the individual items, we gave the credit scoring rule in the larger segment, namely, in the combined polytomous item (Item 2 + Item 3): When the test taker's path went through Node 80, the score was 2; when the path went through Node 83, the score was 1; otherwise, the score was 0.

Moreover, it was noticed that even though a single item in a complex segment had already been designed as a polytomous item, we still could transform the segment into a bigger polytomous item by adding up all scores across items and setting full credit, partial credit, and no credit according to the paths.

3.3. "Bonus" Items

Alternatively, a "bonus" item type could also be present in a complex segment. As the path map shows in Figure 7, students who got a full credit of 1 point in Item 4 (Node 57) had an additional chance to score another point in the subsequent Item 5 (Node 61), while the students who answered incorrectly in Item 4 lost the chance of getting a point in Item 5. The point in Item 5 was a "bonus" for students who gave a correct response in Item 4. Considering that the correlation between Item 4 and Item 5 had a very small chance to be negative, we did not put such "bonus" segments into a polytomous item combination.

Figure 7. *Conversation path of a "bonus item type" example of CPS item with a complex segment in a released CPS unit (The Visit) in PISA 2015.*



Note. The nodes highlighted in yellow are the correct responses to Item 4 and Item 5 respectively.

4. DISCUSSION

Collaborative problem solving is a critical competency in a variety of contexts, including the workplace, school, and home. With the increasing growth of digital tasks, collaborations are not only conducted in real practice but also in the virtual environment. As Dede (2009) has observed, "The nature of collaboration is shifting to a more sophisticated skillset. In addition to collaborating face-to-face with colleagues across a conference table, 21st century workers increasingly accomplish tasks through mediated interactions with peers halfway across the world whom they may never meet face-to-face. … Collaboration is worthy of inclusion as a

21st century skill because the importance of cooperative interpersonal capabilities is higher and the skills involved are more sophisticated than in the prior industrial era."

With the debut of CPS assessment in PISA, it is important to prepare a proper measure to keep the CPS scale reliable and valid. the PISA 2015 CPS units were based on simulated conversations with one or more computer-based agents that were designed to provide a virtual collaborative problem-solving situation. Test takers had to choose an optimal sentence from a multiple-choice list to go through the conversation with agents, or choose one or more actions programmed in the unit. Because of the similar item structures in other domains in PISA 2015, the data collected in the CPS units were evaluated by IRT models (Lord, 1980; Rasch, 1960)—specifically, the two-parameter-logistic model and the generalized partial credit model—to establish reliable, valid, and comparable scales. Readers can refer to the PISA 2015 technical report for the details about scaling and analytic procedures (OECD, 2017). The CPS scale in the main survey consisted of six units, which in turn comprised multiple items within each unit that can be used for the IRT scaling. It was found that data from two units had dependencies in the responses due to different paths that could be taken by students through the simulated chat as a result of the "rescue" item type. Therefore, the CPS chat items that showed this kind of dependency were combined into "composite items" by summing the responses for the different paths that respondents could take. With this approach, it was determined that each path-based response string could be scored to provide valid data and introduced into the IRT analysis. The composite items were used to generate polytomous items for the purpose of reducing issues with local dependencies.

To ensure the computational models were used in an appropriate way, we combined items with high correlations by two steps: first, based on the conversation path analysis, each segment with the "rescue" item type was combined into a polytomous item; and second, the remaining items that still had high correlations in the residual analysis were further combined into a "super" item in the latent trait estimation. This approach is superior in standard large-scale assessments to keep consistent with the whole measurement framework across countries. According to the PISA 2015 tech report, the residuals in CPS domain were under a good control and the local dependency of combined super items were well adjusted.

However, the CPS item design proposed a new challenge in sequential conversation-based measurement. Because of the inherent relationship in conversations, the local independency may not adapt to the assumptions of item response models. Given concerns on the dynamic dependency of at least one previous conservation (or even more), the sequence of the conversation path through the whole unit, i.e., vertical measurement path may be given different difficulty parameters rather than each checkpoint on the conversation line, i.e., horizontal measurement by each item, which the local independency has to be assumed but might not be completely correct.

In addition, the CPS framework with computer agents was compatible with the capabilities of the PISA 2015 computer platform. The student could interact with the agents via a chat window, allowing the student to respond through communication menus. With respect to the student inputs, there were conventional interface components, such as mouse clicks, sliders for manipulating quantitative scales, drag and drop, cut and paste, and typed text input. Aside from communicating messages, the student could also perform actions on other interface components. For instance, additional data could be collected on whether students verified in the CPS environment. These actions were stored in a computer log file, which may provide additional information for tracking students' efforts in solving the CPS units.

Technical advances in computer-based learning systems have made greater efficiency possible by capturing more information about the problem-solving process. Finer-grained information from response time and actions were also added into CPS measurement in recent studies (e.g.,

de Boeck & Scalise, 2020; Han et al., 2023; Qiao et al., 2023). Further, many studies (e.g., von Davier et al., 2019; Han et al., 2019; Gao et al., 2022; He et al., 2021, 2023b; Ulitzsch et al., 2021) showed that process data are more appropriate to describe respondents' behaviors and strategies in interactive tasks. For example, Xiao et al. (2021) applied hidden Markov models on time-stamped action sequence data to identify the latent states and transitions between states underlying the problem-solving process. Ulitzsch et al. (2023) explored the early predict-ability of behavioral outcomes on interactive tasks with early-window clickstream data. They applied extreme gradient boosting to dynamically classify respondents who have a high probability of being out of track when solving a problem-solving task. He et al. (2023a) developed dynamic time warping method to cluster students' dynamic navigation patterns. These methods are worth further exploration to investigate the associations between sequences of actions and CPS skills and to extract sequence patterns for different CPS proficiency levels.

Considering the complexity of human-to-human interaction in collaborative conversations across countries and languages, PISA 2015 adopted the human-agent module in CPS domain. This new item type also brings challenges in test translation and fairness across countries in diversified cultural environments. It would be interesting to check for test fairness across different language groups and investigate the effect of languages in the CPS measures in a future study. The advances in text-based generative artificial intelligence applied in large language model (LLM; OpenAI, 2023) shed lights on alternative approaches to handle conversation-based assessment in the near future, which might be self-trained on different languages.

In conclusion, PISA 2015 CPS competency is a conjoint dimension of collaboration skills, functioning as a leading strand, and problem-solving skills, functioning as an essential perspective. The effectiveness of CPS depends on the ability of group members to collaborate and prioritize the success of the group over that of the individual. At the same time, this ability is a trait in each of the individual members of the group (OECD, 2013). The methods in PISA 2015 introduced in this study for collaborative item design could be applied to other collaborative human-agent items in similar settings and also challenge other researchers to refine the methodology and add extra information or data sources to get a better CPS scale. For future studies, we recommend using multivariate statistical analyses to address different aspects of CPS units and combining these analyses with process data from log files to track the process of students' learning and collaborative activities.

Acknowledgments

The author thanks Mary Louise Lennon, Henry Chen, Matthias von Davier and Hyo Jeon Shin for helpful suggestions at the initial stage of this study; the Center for Global Assessment at Educational Testing Service (ETS) for their support. All views expressed in this paper are solely those of the author and do not necessarily reflect those of the OECD or ETS.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

Orcid

Qiwei He https://orcid.org/0000-0001-8942-2047

REFERENCES

Casner-Lotto, J., & Barrington, L. (2006). *Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. workforce*. http://www.conference-board.org/pdf_free/BED-06-Workforce.pdf

Cooke, N.J., Duchon, A., Gorman, J.C., Keyton, J., & Miller, A. (2012). Preface to the special section on methods for the analysis of communication. *Human Factors: Journal of the Human Factors and Ergonomics Society, 54*, 485–488.

de Boeck, P., & Scalise, K. (2019). Collaborative problem solving: Processing actions, time, and performance. *Frontiers in Psychology, 10*, 1280.

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science*, *323*(5910), 66-69.

Dede, C. (2012). *Interweaving assessments into immersive authentic simulations: Design strategies for diagnostic and instructional insights*. Paper presented at the Invitational Research Symposium on Technology Enhanced Assessments. http://www.k12center.org/rsc/pdf/session4-dede-paper-tea2012.pdf

Gao, Y., Cui, Y., Bulut, O., Zhai, X., & Chen, F. (2022). Examining adults' web navigation patterns in multi-layered hypertext environments. *Computers in Human Behavior, 129*, 107142.

Greiff, S., Wüstenberg, S., Csapo, B., Demetriou, A., Hautamäki, J., Graesser, A. C., & Martin, R. (2014). Domain-general problem-solving skills and education in the 21st century. *Educational Research Review, 13*, 74-83.

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

Han, Z., He, Q., & von Davier, M. (2019). Predictive feature generation and selection using process data from PISA interactive problem-solving items: An application of random forests. *Frontiers in Psychology, 10*, 1421.

Han, A., Krieger, F., Borgonovi, F., & Greiff, S. (2023). Behavioral patterns in collaborative problem solving: a latent profile analysis based on response times and actions in PISA 2015. *Large-scale Assessments in Education*, *11*(1), 35.

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Identifying generalized behavioral patterns with sequence mining. *Computers & Education*, *166*, 104170.

He, Q., Borgonovi, F., Suárez-Álvarez, J. (2023a). Clustering Sequential Navigation Patterns in Multiple-Source Reading Tasks with Dynamic Time Warping Method. *Journal of Computer-Assisted Learning*, *39*(3), 719-736.

He, Q., Shi, Q., Tighe, E. (2023b). Predicting problem-solving proficiency with hierarchical supervised models on response process. *Psychological Test and Assessment Modeling*, *65*(1), 145-178.

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 749-776). Information Science Reference.

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). Springer.

Hirschberg, D.S. (1975). A linear space algorithm for computing maximal common subsequences. *Communications of the ACM*, *18*, 341-343.

Hirschberg, D.S. (1977). Algorithms for the longest common subsequence problem. *Journal of the ACM*, *24*(4), 664-675.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.

National Academies. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. http://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_070895.pdf

National Center for Education Statistics (2015). *The nation's report card: 2015 mathematics and reading assessments.* Publication No. NCES 2015136. Washington, DC: Author.

OpenAI. (2023). ChatGPT (May 24 version) [Large language model]. https://chat.openai.com/chat/

Organisation for Economic Co-operation and Development (2013). *PISA 2015: Draft collaborative problem solving framework*. Paris, France: Author.

Organisation for Economic Co-operation and Development (2015a). *PISA 2015 released field trial cognitive items.* Paris, France: Author.

Organisation for Economic Co-operation and Development (2015b). *PISA 2015 field trial analysis report: Outcomes of the cognitive assessment (JT03371930).* Paris, France: Author.

Organisation for Economic Co-operation and Development (2017). *PISA 2015 Results (Volume V): Collaborative Problem Solving*, PISA, OECD Publishing, Paris, France.

Qiao, X., Jiao, H. & He, Q. (2023). Multiple-Group Joint Modeling of Item Responses, Response Times, and Action Counts with the Conway-Maxwell-Poisson Distribution. *Journal of Educational Measurement, 60*(2), 255-281.

Quellmalz, E.S., Timms, M.J., & Schneider, S.A. (2009). *Assessment of student learning in science simulations and games*. Paper prepared for the National Research Council Workshop on Gaming and Simulations, Washington, DC.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rosenbaum, P.R. (1988). Item bundles. *Psychometrika*, *53*(3), 349-359.

Stecher, B.M., & Hamilton, L.S. (2014). *Measuring hard-to-measure student competencies: A research and development plan,* Research Report. RAND Corporation.

Ulitzsch, E., He, Q., Ulitzsch, V., Nichterlein, A., Molter, H., Niedermeier, R., & Pohl, S. (2021). Combining clickstream analyses and graph-modeled data clustering for identifying common response processes. *Psychometrika, 86*, 190-214.

Ulitzsch, E., Ulitzsch, V., He, Q., & Lüdtke, O. (2023). A machine learning-based procedure for leveraging clickstream data to investigate early predictability of failure on interactive tasks. *Behavior Research Methods*, *55*, 1392–1412.

von Davier, M., Khorramdel, L., He, Q., Shin, H., & Chen, H. (2019). Developments in psychometric population models for data from innovative items. *Journal of Educational and Behavioral Statistics, 44*(6), 671-705.

Wildman, J.L., Thayer, A.L., Pavlas, D., Salas, E., Stewart, J.E., & Howse, W. (2012). Team knowledge research: Emerging trends and critical needs. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 54*, 84-111.

Wilson, M., & Adams, R.J. (1995). Rasch models for item bundles. *Psychometrika*, *60*(2), 181-198.

Xiao, Y., He, Q., Veldkamp, B.P., & Liu, H. (2021). Exploring Latent States of Problem-Solving Competence Using Hidden Markov Modeling on Process Data. *Journal of Computer-Assisted Learning, 37*(5), 1232-1247.