Comparative analysis of student performance in collaborative problem solving: What does it tell us?

Background

Collaboration skills have been increasingly identified as important for success in school and work environments (O'Neil et al., 2004; Singh-Gupta & Troutt-Ervin, 1996). As a result, educational research on collaboration has been in abundance in recent years (OECD, 2017; Griffin & Care, 2015; von Davier & Halpin, 2013; Scoular & Care, 2019). In particular, the decision to measure collaboration in a problem solving context has been a particular focus, in order to provide context and purpose to the application of collaborative skills. The OECD's decision to assess collaborative problem solving (CPS) in the Programme for International Student Assessment (PISA) in 2015 has been a major driver in highlighting the importance of measuring such skills.

In education systems around the world, teachers are being tasked with monitoring and improving students' collaboration skills (Scoular et al., 2020). One of the major challenges in that endeavour is identifying exactly what collaboration looks like in the classroom, and how student proficiency in it can be described. The PISA-CPS assessment measured collaboration one specific way, by having students interact with computer agents - who were not real people but sets of programmed responses or prompts - to solve problems. This paper explores what the data from this assessment tells us about collaborative problem solving, and how this data compares with data from two other assessments of collaboration. Evaluation can be made about what assessment features are required and working well for developing measures of this innovative domain.

Measuring collaboration in PISA

Alongside the assessment of the traditional domains of science, mathematics and reading, OECD's Programme for International Student Assessment (PISA) introduced the assessment of the innovative domain of collaborative problem solving (CPS) for the first time in 2015. About 125,000 15-year-old students in 52 countries and economies participated in the computer-based CPS assessment.

PISA measures individual competency and, in the domain of CPS, it measures the ability of individuals to work in collaborative settings. To achieve this, in PISA-CPS, the pupils interacted with computer agents instead of other humans in a computer-based assessment. The assessment was developed to measure the CPS skills over various computer-simulated assessment tasks. Each task involved a scenario with multiple individual items that students had to work through. In order to communicate with other group members (i.e. computer agents), students had to select a response from a list of predefined messages displayed at the task space. Actions such as clicking or dragging and dropping were implemented in the task space. Each correct action or message reflected a specific CPS skill.

Students in Australia performed higher than the OECD average in CPS (531 points compared with 500 points), meaning their ability to achieve successful outcomes in collaborative settings was higher than that of an average 15-year old across OECD countries. In Australia, girls scored 41 points higher than boys on average in CPS (552 points compared with 511 points). Regarding the traditional learning areas, students in Australia had an average of 510 points in science, 503 points in reading and 494 points in mathematics. Their performance in CPS was highly correlated with the three learning areas (r=.76 with science, r=.75 with reading and r=.68 with mathematics) (Avvisati & Keslair, 2014). These findings are crucial in understanding a complex skill such as CPS and how it relates to other learning areas.

Collaboration framework

The definition of collaboration is much more complex than simply working with others. The literature has shifted from a simple definition of collaboration as working in groups, to defining collaboration as an action where two or more learners pool knowledge, resources and expertise from different sources in order to reach a common goal. The distinction between interdependence and independence provides some insight into the nature of collaboration. While the focus of team or group work literature has been on independent teams where learners work in relative isolation, interdependent teams rely on the actions of others and cannot perform activities independently (von Davier & Halpin, 2013). Collaboration is related to the latter. There is shared responsibility and an active division of labour.

Although there are different definitions of collaboration presented in the literature, similar components can be identified in each (OECD, 2017; Hesse et al., 2015). For example, due to the nature of collaboration, the participation of each learner and their level of engagement with an activity directly impacts on the effectiveness of the collaborative group as a whole. PISA 2015 defines CPS competency as: “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, 2017b, p.134). The PISA-CPS framework identifies two components of CPS: the four cognitive processes identified for individual problem solving in PISA 2012, and three competencies unique to CPS. These three competencies are crossed with the four problem solving processes to form a matrix of twelve specific skills, as illustrated in Table 1 above. This definition was designed for a specific purpose – to provide cross-country measures of students’ collaboration in problem solving contexts when working with computer agents.

INSERT TABLE 1 HERE

The results of the PISA-CPS assessment, alongside results from two other major assessments of collaboration (Scoular, 2017; ACER, 2020) were instrumental in generating a better understanding of the skill and providing sufficient evidence of the skill. In an effort to adopt a broader sense of collaboration, ACER developed a framework taking into consideration the aforementioned assessments as well as relevant literature to present a synthesised definition of collaboration. The ACER Framework for Collaboration (Scoular et al., 2020) was developed from these assessments and their associated frameworks as well as many other existing definitions of collaboration (e.g. von Davier & Halpin, 2013; Dillenbourg 1999; O'Neil et al., 2004) and later validated by the assessment data to provide a suitably detailed and comprehensive framework for describing collaboration. In this study we use this extended framework as a common reference for mapping items across assessments, including released items from PISA-CPS, to demonstrate and develop a comprehensive understanding of the skill.

The purpose of the ACER framework is to establish a common terminology for describing collaboration, taking into consideration all major assessments of collaboration to date, and providing a structure that is suited for the assessment and teaching of collaboration. The framework describes collaboration within strands that are then further qualified as aspects (see Figure 1). Specifically, a strand refers to the overarching conceptual category for framing the skills and knowledge addressed by collaboration assessments, while an aspect refers to the specific content category within a strand.

INSERT FIGURE 1 HERE

Assessments

Measuring an innovative domain requires innovative measures. Evidence of ability in such skills is likely to be covert, not directly observable, and therefore, inferences about student ability need to be drawn from demonstrated behaviours. The drawing of such inferences can be facilitated by the use of technology in assessments. Three assessments, including PISA-CPS, were selected for comparison in this study. These specific assessments were selected as the data was available to the authors, they all utilised technology for test delivery, and their focus was on collaboration in problem solving contexts. Further, selection had the the aim of covering a diverse range of assessment designs, different to the one adopted by PISA-CPS. For instance, PISA-CPS used computer agents as team members in the collaborative setting, while the other assessments reviewed in this paper used real humans. Another difference lies in the communication, while PISA-CPS used a restricted environment of multiple-choice messages to facilitate communication, the other two assessments included a chat box where students could type their responses. Psychometric analyses with the aim of examining item quality were performed for each of the three assessments and to further demonstrate how assessment can support understanding of a complex skillset.

The first assessment under review is the CPS assessment administered as the innovative domain of PISA-CPS. The second assessment called Delivery Hero (DHO-CPS) was developed by Scoular (2019) and was intended as an advancement of the Assessing and Teaching of 21st century skills project, an international study focusing on developing formative assessments and resources for developing CPS and ICT digital literacy skills (ATC21S; Care et al., 2016). The third assessment was developed by ACER to measure 'general capabilities', one of which was collaboration (ACER-GC). Table 2 summarises the features of each assessment (for full details see PISA-CPS: OECD, 2013; DHO-CPS: Scoular, 2019; ACER-GC: Scoular & Timms, 2020). All three assessments measured aspects of collaboration within problem solving based scenarios. The data pertaining to Australian students is focused upon in this study.

INSERT TABLE 2 HERE

Measures

The PISA-CPS assessment comprises 117 activities distributed across 6 scenarios. Student actions and responses were captured and scored based on the PISA framework (OECD, 2017b). Each item was coded in two (dichotomous: 0/1) or more (polytomous: 0, 1, ... m) categories according to the item coding rubrics. That is, depending on students' actions and responses they received credit, partial credit or no credit. Further details with respect to the PISA-CPS test design can be found in the technical reports of the study (OECD, 2017a, 2017c). PISA-CPS uses several different types of CPS tasks (e.g., jigsaw or hidden-profile, consensus-building, and negotiation tasks).

Figure 2 illustrates a snapshot of the released PISA-CPS task named "Xandar". In this task, a three-person team consisting of the student test-taker and two computer agents (Alice and Zach) take part in a simulation of an in-class contest where they must answer questions about the geography, people, and economy of the fictional country of Xandar. The unit involves decision-making coordination tasks and requires consensus-building collaboration (OECD, 2017c). The sample item illustrated in Figure 2 requires students to help team members negotiate a solution when conflict arises. In this case both Alice and Zach express their interest in answering questions about the people of Xandar. The credited response to this item solicits additional information about each team member's point of view (OECD, 2017c).

INSERT FIGURE 2 HERE

DHO-CPS. As students worked through solving nine collaborative problem solving scenarios, behaviours including actions and chats were recorded in log files. Where a behaviour was found in the log file a score of '1' was allocated to that student; or '0' where a behaviour was absent. This resulted in scores for each student on each behaviour in each activity resulting in 72 items.

ACER-GC. The assessment comprises 11 activities based around a single scenario. It uses a combination of HTML pages and software from Google to host and deliver the test content to students via a web-browser. Google documents contain activity instructions and tables in which students can enter information as the response format. Google Hangouts is used to host chats between group members where students collaborate on activities.

Analysis part 1: Mapping of assessments to the ACER Framework for Collaboration

Each of the three assessments' items were reviewed and mapped to one of the aspects in the ACER Framework for Collaboration. Table 3 below presents the number of items in each assessment mapped to each aspect (for full details of each aspect see Scoular et al., 2020).

INSERT TABLE 3 HERE

Following the mapping of items from the three assessments to the aspects of ACER collaboration framework, it is evident that none of the assessments include items pertaining to the aspects of "Adapts behaviour and contributions for others". In addition, two of the assessments (i.e., PISA-CPS and DHO-CPS) did not cover the aspect "Negotiates roles and responsibilities".

For the remaining aspects, the PISA-CPS assessment covered all of them with at least one item, with the exception of the aspect "Pools resources and information" that was not covered at all by this assessment. Similarly, DHO-CPS covered all the remaining aspects with at least one item. It is worth noticing that, in the DHO-CPS assessment, the aspects "Communicates with others" and "Maintains shared understanding" have a disproportionately high number of items representing them. Finally, the ACER-GC assessment covered all remaining aspects with at least one item, apart from the aspect "Engages with role and responsibilities", which was not represented.

Possible reasons for a lack of mapping to some aspects are explored in the discussion section. One clear message that comes out of our mapping exercise is that none of the three assessments covered all hypothesised aspects of the collaboration framework used for mapping.

Analysis part 2: Item Response Theory analysis

Each set of assessment data was analysed using Item Response Theory (IRT), specifically a 1 Parameter Logistic model, or Partial Credit model where items are polytomous (Rasch, 1960; Wright & Masters, 1982). These models identify the probability of a student giving a correct answer to an item as a function of that student's underlying ability on the latent trait and the difficulty of that item. Calibration of PISA-CPS included 117 items, of which 20 were coded as partial credit and 97 as dichotomous. All 72 items in DHO-CPS were coded as dichotomous. ACER-GC contained 15 items, 3 partial credit and 12 dichotomous. The data was modelled using ACER ConQuest version 5 (Adams et al., 2020) and secondary analysis and visualisations were produced conquestr (Cloney & Adams, 2020) in R (R Core Team, 2019).

Each model was estimated as a single dimensional model, with items scored as increasing integers (zero for incorrect, and increasing integers representing increasing correctness for each item). The 1PL model is fit to the data, using Marginal Maximum Likelihood (MML) estimation. All models converged using standard criteria for the expectation–maximization (EM) algorithm.

To consider fit of the model to the data weighted (sometimes called Infit) and unweighted (sometimes called Outfit) fit are examined as evidence that the underpinning construct was represented by the indicators. Item fit statistics indicate how accurately the model fits the data, through an analysis of the residuals, under the assumption of the 1PL or PCM, including equal discrimination, model misspecification, and unidimensional. Rules of thumb have been developed to describe acceptable fit, this study adopts a range of 0.70 and 1.30 (Wu, 1997; Adams & Wu. 2009). All items in each assessment fell well within this range (see Figure 3 for DHO-CPS items) suggesting that the set of items fit the underlying model well and there is an assumed single unidimensional collaboration construct being measured by each of the assessments.

INSERT FIGURE 3 HERE

The separation reliability of the unidimensional collaboration scales was high. The item separation reliability was 0.99 for each assessment. The expected a posteriori/plausible value (EAP/PV) reliability was also high at 0.90 for PISA-CPS, and 0.90 for DHO-CPS, and somewhat lower at 0.68 for ACER-GC. [1] These indices indicate that items on the continuum of each scale are well separated and the three assessments are each sensitive enough to differentiate well between student collaborative abilities.

[1] The lower reliability for the ACER-GC scale is to be expected given there were fewer items overall and, relative to this, it is considered a moderate-high reliability.

The importance of model fit can be visualised through item characteristic curve plots. Smooth lines represent the model expectation of the change in the probability of endorsing that response category (y-axis) as latent ability increases (x axis). Figure 4 illustrates a good fitting dichotomous item (CC104101) and an example of a poorer fitting partial credit item (CC102209C) from the PISA-CPS scale. Misfit is seen in departure from the smooth line by the joined dots of the empirical line – this is the data that has been observed. For the poorer fitting item, it can be seen that misfit is particularly present at the low end of the ability continuum and the misfit can be described as under-discrimination as the empirical line is flatter (and therefore fit statistic is greater than 1: 1.28). This item measured “Participates in the group” in PISA-CPS and identifies whether the student enacted the intended plan. Given that enacting the intended plan is explicitly instructed as part of the activity, this feature of test design could have reduced the discrimination between lower and higher ability students, resulting in a slightly flatter line than expected. While still in the acceptable range, this item could be further reviewed to tell us more about the construct and the impact of test design on the item quality.

INSERT FIGURE 4 HERE

In addition to item fit, an important step is to assess targeting. That is, the item difficulty should not be so difficult or so easy that no student could reasonably answer correctly (or that all students will answer correctly). An item-person map (sometimes called a Wright Map) visualises the location of the item difficulty, either by plotting the category (Thurstonian) thresholds or the item deltas (difficulties). In the case of dichotomous items, these are the same value. Figure 5 shows an item-person map for the PISA-CPS scale, using item deltas. For polytomous items, this is the mean of the item threshold parameters, called taus in the delta plus tau specification (Andrich, 1978), which should not be confused for Thurstonian thresholds. The curve shows the distribution of student abilities (see the peak, approximate mean at 0.5) and the location of the item difficulties. There is good coverage of items along the scale, with most items (darkest grey, where items overlap) around the middle of the scale. This is consistent with the other two scales that show good targeting.

INSERT FIGURE 5 HERE

Taken together, this validation work shows that the three scales are functioning reasonably well, demonstrating good fit to the model, good reliability, and good targeting. Note that much analysis has been omitted for the sake of brevity. Given the models presented, we conclude that each of the assessments are measuring a single construct, and we interpret this construct to be collaboration in the context of problem solving.

Discussion: What have we learned about assessing collaboration well?

The comparison of the three assessments in this study have provided insight into assessing collaboration. Numerous attempts have been made to elucidate the principles of good test development (see for example Mendelovits, 2017). For example, all items should be as transparent as possible, so that the challenge for students is in responding to the stimulus material, and the wording of the items does not pose an extraneous comprehension load that would lead to construct-irrelevant variance in responses. This means the item should be succinct, should have no tricks, ambiguity or difficult language and should avoid negatives. It is not difficult to apply such general principles to assessments of collaboration. There are, however, areas that pose a particular challenge to test developers in relation to the assessment of collaboration.

Coverage of the construct

There is good construct representation in all three assessments as their items are distributed across the three strands of the framework. However, there are two aspects that are not well represented across assessments: “Negotiates roles and responsibilities” and “Adapts behaviour and contributions for others”. Importantly, despite the lack of items assessing these aspects, they remain a part of the construct of collaboration - research literature strongly points to the importance of both (Scoular et al., 2020). What the lack of items assessing these constructs suggests, rather, is that despite the centrality of these aspects, test developers have had limited success in eliciting these skills in an assessment context to date. Exploring ways to elicit evidence of these aspects in future assessments would assist in developing a greater understanding of how these aspects are demonstrated by students, and therefore how to identify evidence of them more readily.

“Adapts behaviour and contributions for others” is not measured in any of the three assessments. This aspect relates to students identifying an appropriate style and level of complexity relevant to their group members and being able to adjust their communication, behaviour, and contributions to suit other group members’ needs. Commonly referred to in the literature as receiver awareness, it is a valuable skill for coordinating mutual activities (Dehler et al., 2011). Proficient collaborators tailor their behaviours and contributions to suit others based on their interpretation of their peers’ understanding. This is anticipated to be a very difficult behaviour to capture in the classroom, and more so in computer-based delivery. Therefore, while still an important aspect in the collaboration framework, it is understandable that this aspect has so far not been a major focus of assessment. Evidence of this aspect may be accessible for test developers to elicit – for example, through activities simply asking students to identify which form of words would be most appropriate for a given audience.

“Negotiates roles and responsibilities” was only measured in ACER-GC. This assessment specifically tailored an activity to focus on this aspect, whereas, the other two assessments had pre-determined roles for each student, and thus no roles to negotiate. Therefore, the PISA-CPS and DHO-CPS assessments were not designed to elicit evidence in relation to this aspect. The definition of collaboration outlined in an earlier section makes clear that it is about shared responsibility and an active division of labour. The measurement of this aspect then is dependent on the design of the collaborative activity – the opportunity to select and allocate roles according to the different tasks that must be completed in order to achieve the common goal.

“Pools resources and information” is only frequently measured in DHO-CPS. This aspect is targeted by items in which students are expected to share resources or information when their peers ask them a question. This is identified in DHO-CPS when students pass resources to another group member’s screen or share their own screen view with others. These functionalities are not available in the other two assessments. It could be concluded that this aspect is highly dependent on the nature of resources or information that the student is expected to pass to their peers. When this is something simple like dragging and dropping a box, the aspect is easy to measure, but when evaluation of the information is needed through chat logs, then the aspect is more difficult to measure. Assessments can be designed to facilitate this by providing a space in which students can opt to pool their resources, and the contributors adding to this space can be monitored and assessed.

Test design

Activity length

In order to capture the richness of collaboration, complex activities need to be built. This complexity typically requires many activities and processes to be captured, which translates into a great deal of student time on the assessment. Each of the three assessments evaluated present multiple activities of various lengths. There is a trade-off here. Shorter activities may be preferable so as to avoid overburdening students, but shorter activities are unlikely to allow students to demonstrate the full range of their skills. One way to address this issue is by designing sets of activities that try to strike a balance between adequately sampling the different aspects so as to allow an estimate student’s collaborative ability, while still attempting to minimise the overall testing burden.

Distribution of resources

All three assessments provide symmetry in regards to goals (that is, all students within a group are working toward the same goal) but they present opposing perspectives on symmetry of the resources. The nature of collaboration relies upon an asymmetry of resources between students, representing a real-world view of diversity of expertise, knowledge and information (Scoular at al., 2017). Therefore, assessments in which students possess different resources are likely to allow better and more authentic coverage of the construct. DHO-CPS activities vary in the degree of asymmetry with most presenting varying perspectives of the problem solving scenario and distribution of resources between students. This set up should encourage students to work collaboratively as without one another they do not possess enough information or resources to solve the problem. The resource asymmetry presented within an activity leads to an individual asymmetry between students as each individual bring the resources to bear in a different way. By adopting a Human-Agent approach, the PISA activities remove any possible asymmetry within activity or between individuals since there is only a single student.

Group composition

Collaboration, by definition, requires the formation of a group. How such groups should be composed is not a trivial question. Individuals may perform differently, depending on the group to which they are assigned, with factors such as differences in ability (Wildman et al., 2012), personality characteristics (McGivney et al., 2008) and gender (Bear & Wooley, 2011) all potentially influencing how a group might collaborate. In the case of PISA-CPS, no consideration was needed in relation to group composition, since computer-agents were used as team members, pre-programmed to cover a variety of behaviours. In DHO-CPS, students were randomly teamed together, and in ACER-GC, group formation was based on teacher decision.

An overarching question in relation to group composition is whether all members of a group should be human. Computer delivery not only makes it possible for students in different geographical locations to collaborate, but also allows collaboration between a student and one or more virtual agents. In PISA-CPS, students worked not with other students, but with computer agents (OECD, 2017). This approach has the dual advantages of controlling for group dynamics and simplifying the process of data capture, since there would be no need to capture real-time interactions between group members. On the other hand, the use of computer agents rather than real students to assess collaboration has been viewed as lacking in authenticity (Scoular et al., 2017).

The issue of how groups should be formed is an important one, since group composition has the potential to either enhance or suppress an individual’s ability to show their own skills. In practice, how this issue is dealt with is likely to depend on the nature and stakes of an individual assessment. In a high stakes assessment with strict requirements for standardisation, it will be necessary to pre-designate groups in some way. By contrast in lower-stakes classroom activities, the most pragmatic approach might be to allow the teacher to form groups based on what is likely to ease the classroom management of the activity, or for pedagogical considerations. Decisions about group composition for any assessment of collaboration should be made with the purpose and the nature of the evidence it aims to collect firmly in mind, with the knowledge that such decisions may influence the aspects of the construct that can be elicited by the assessment.

Capturing communication

Process data is a useful tool in educational assessment and, is particularly useful for modelling and evaluating collaborative assessments (Adams et al., 2015; von Davier & Halpin, 2013; Zoanetti, 2010). In particular the contents, sequence, and frequency of communication between students in the process data stream can provide detailed indicators of the dynamics between students.

Capturing communication while students are collaboratively solving problems can provide an abundance of rich data (Okada & Simon, 1997; Palincsar & Magnusson, 2001; Teasley, 1995). The communication demands of collaboration may provide an insight into cognitive processes that may otherwise not be accessible. However, analysing and interpreting the chat content is difficult. Quantitative metrics such as count of actions and measures of time are useful, although they do not provide the full picture of the nature of what is being measured. Log file analysis focusing on the frequency of occurrences of chat is popular in the collaborative learning literature. However, the quantity of actions reveals little regarding the quality of the communication. More frequent communication does not necessarily indicate better collaboration; in fact, it could indicate inefficient communications (Meier et al., 2007).

A study using open chat boxes to gather the communication between students, as in the DHO-CPS assessment, found that the syntax and grammar was too inaccurate to analyse automatically (Adams et al., 2015). For example, many students send chat in chunks, not full sentences, and many students use abbreviations in their communications which would not be accounted for by such programs. However, keywords and placement of communications in relation to actions are able to be captured and can provide insight into student activities in the assessments.

In the ACER-GC assessment, the communication is hand-scored by expert markers using scoring rubrics that target different aspects of collaboration. Other options, perhaps using a combination of human and machine scoring of responses might also be possible. It is the case though, that scoring of human to human collaboration is likely to be labour intensive, and this is in part a reflection of the fact that collaboration is a complex skill, one that may not be possible to assess and code simply. By comparison, in the PISA-CPS assessment students choose from a list of pre-defined responses. While neither approach is completely effective, an optimal approach may be a combination of the two, wherein chat communication between students in the open chat boxes, as in DHO-CPS, is used to develop response options. Although the response options will still be constrained, these will be based on real student communication in the same scenarios and will allow for automated scoring.

In summary, interpretation of communication in process data can be difficult but technology is becoming increasingly refined and ongoing research will point to new innovations in this space. When assessing collaboration, the placement of the chat between significant actions could be of more value than frequency of chat, to interpret the quality of the interaction.

Features of collaborative tasks

In attempting to measure collaboration, as defined in the ACER framework, there appear to be some essential features required of a collaborative assessment:

- A problem to be solved (preferably ill-defined in nature)
- Resources that students can move around within, between, and across environments
- Division of resources and information between students
- A chat box – as a communication device

Collaborative activities must present a scenario that necessitates that two or more collaborators work together, else there is little purpose to the collaboration (Scoular, 2019). A problem to be solved also provides context and purpose to the collaboration. If students can solve the problem by themselves there is little motivation for the students to work with others. There has to be a requirement for collaborating beyond merely a division of labour. This should be managed through different resources and information being made available to the different collaborators with each student having their unique set of resources, resources which students can manipulate on their own screen, as well as resources which they can pass between screens to their peers. This allows for student interactions, with their resources and partner, to be monitored and interpreted. It also necessitates that each student contributes to the activity, providing a platform in which students can demonstrate their collaborative skills.

Another requirement based upon the definition of collaboration is an opportunity to communicate with others. Communication must occur in a way that can be recorded and interpreted using the assessment framework. A chat box provides a means to communicate and the format of that means is critical to how well students can interact and feedback to one another (Zagal et al., 2006).

Scoring: What should be scored?

Group vs individual

When assessing something that is collaborative, an immediate question is whether one should assess each individual within a group, the group as a whole, or both. Of the assessments examined in this paper, PISA assessed individuals only (a consequence of the decision to use human-agent collaboration). Both DHO-CPS and ACER-GC contained some activities that assessed individuals, and others in which a group score was given. While this may seem an ideal solution, it presents its own challenges in relation to the choice of statistical model. IRT models, for example, assume independence. If group scores are given, all students in a group receive the same score on a particular item, and this assumption is violated. This observation reinforces the point that assessing skills may require new approaches to both assessment, and scoring.

Product vs process

A related, but different issue of scoring is whether, given the interactive nature of collaboration, one should assess some final group product, some part of the collaborative process, or both. From a purely practical point of view, both are possible, particularly when utilising computer-based assessment. When an assessment is delivered via computer, every interaction of the test-taker with the test may be easily recorded, making it possible to collect far more information about student activities during the test than has previously been possible (Baker & Mayer, 1999; Chung & Baker, 2003; Schacter et al., 1999). While it is now possible to record such information, what must be addressed if one is to consider making use of such data is whether it is possible to use it to derive valid measures of the construct (Ramalingam & Adams, 2017). That is, to what extent do each of the final product, and measures of the process, reflect the construct of collaboration? It may be that valid measures of collaboration can be derived from both a final product, and well-defined aspects of the collaborative process. In the ATC21S project (Griffin et al., 2012), on which DHO-CPS drew, it was assumed that the collaborative process should be valued (and assessed) in its own right (that is, independent of a final solution), and parts of the collaborative process were targeted within the assessment. Whether to assess a final product, or parts of a process should be explored in relation to all new assessments, always ensuring that if the latter approach is taken, that measures of the process can validly act as evidence of skill in the construct being assessed.

In the context of the three assessments, it has been possible to identify behaviours associated with student exploration of the assessment environment for the task, development of their understanding of the problem, joint planning, contributions to solutions and their evaluations of progress. Identification of sequences of actions and interactions in log stream data provides insight into the processes students undertake. These processes allow a better understanding of the construct of collaboration than a mere solution could allow and understanding the processes of collaboration is likely to be more valuable for teaching intervention. An educational assessment should not stand in isolation but must be aligned with curriculum and instruction if it is to support learning (Pellegrino et al., 2001). Being able to identify at which part in the collaborative process a student struggled or passed through can help educators to make decisions regarding the instructional needs of a student (Zoanetti, 2010). Measuring the solution and whether it is correct is undoubtedly a useful criterion but not as stand-alone evidence of student collaborative ability. Instead, a set of indicators could help identify which steps each student went through to gain that outcome. Teachers would benefit from this information when determining how best to improve their students' collaborative ability.

Issues of measurement

Human-Computer Agent vs. Human-Human

One of the limitations in PISA-CPS is that students interacted with computer agents rather than other students through face-to-face interactions or computer-mediated communication. The use of computer agents satisfied various logistical challenges, but it still raised the concern that the assessment environment deviated from naturalistic, ecologically valid collaborative activities (Graesser et al., 2018). Although OECD (OECD, 2017c) reports that PISA-CPS is informative about students' performance in real-life collaboration scenarios, the differential impact of computer agent and human impact on student responses remains unresolved (Scoular & Care, 2019). Research work investigating whether computer agents can validly replace humans found no significant differences between the types of collaboration partner (Herborn et al., 2020). However, the generalisability of the results is limited, since significant constraints on the human-human collaboration were still posed (i.e., free chat response was prohibited, while predefined chat communication was adopted). In a similar vein, researchers trying to apply the PISA assessment approach in their studies highlighted that they found it hard to design a problem scenario assessing skills reflecting establishing and maintaining team organisation, mostly because it was difficult to always design a response format for the computer agent that was sufficiently "human-like" (Lin et al., 2015).

A second limitation of using computer agents is that, in assessments like PISA-CPS, the actions of the individual student were constrained to a small set of choices to allow researchers to associate each action with a particular collaborative skill. Although using lists of pre-defined messages has been argued to provide a tractable way of measuring communication and collaboration skills (Hsieh & O'Neil, 2002), constraints imposed to the number of possible discourse patterns that students could perform, posed challenges in allowing for and eliciting students' creativity, their ability to introduce new ideas, and negotiate or alter previous actions/messages. For instance, negotiation, a very important conversational pattern that is part of establishing shared knowledge, making a decision, or agreeing on a course of action, often takes a multi-turn exchange between team members to happen, but PISA-CPS allowed only one exchange rather than multiple exchanges (Graesser et al., 2018). This limitation was also confirmed from the mapping of the existing and available assessments onto the ACER collaboration framework showing that they have been unsuccessful in eliciting behaviours reflecting negotiation. Finally, as pointed out by Çakır et al. (2009), such an approach unduly restricts interaction which must be flexible enough to allow students to engage in unanticipated behaviours.

Conclusion

It is increasingly apparent that our understanding of complex skills needs to be rapidly developed in order to meet the demand of 21st century education (OECD, 2017; Griffin & Care, 2015; von Davier &

Halpin, 2013; Scoular & Care, 2019). Complex skills such as collaboration and problem solving can be difficult to teach and learn but the analysis from this study demonstrates that robust measurements can be developed that provide insight into how these skills can be demonstrated. Further, assessment data can actually provide more information and improve understanding of such complex skills. The assessments explored in this study have led to increased understanding about what behaviours and processes can be associated with aspects of collaboration. Interpretations of the data visualisations such as the Wright maps can indicate how different proficiencies of collaboration might be demonstrated. The definition and measurement of CPS has already evolved in the short period of time since PISA in 2015. The development of other assessments since, and the use of different technologies has provided different types of information about the skill that has in turn informed the definition and understanding of it. The comparison of these three different assessments of collaboration also indicates that measuring the skill is of importance for different purposes: PISA-CPS being large scale, comparative and policy orientated, and the other two assessments being classroom and formative assessment orientated.

This study, through the comparison of three different assessments in collaborative problem solving, contributes valuable research in developing best practices in measuring collaboration, particularly in relation to the considerations covered such as test design, scoring criteria, and representation of the construct within assessments but in broader terms support ideas of teaching and learning such complex skills. Assessment of such skills, particularly in relation to growth, can shed light on how to appropriately situated teaching interventions and to identify learning in an innovative domain.

References

Adams, R. J., & Wu, M. (2009). The Construction and Implementation of User-Defined Fit Tests for Use with Marginal Maximum Likelihood Estimation and Generalized Item Response Models. *Journal of Applied Measurement, 10*(4).

Adams, R., Vista, A., Scoular, C., Awwal, N., Griffin, P., & Care, E. (2015). Automatic coding procedures for collaborative problem solving. In P. Griffin and E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 115-132). Springer. https://doi.org/10.1007/978-94-017-9395-7_6

Adams, R. J., Wu, M. L., Cloney, D., & Wilson, M. (2020). *ACER ConQuest: Generalised Item Response Modelling Software (Version 5)* [Computer software]. Australian Council for Educational Research.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573. https://doi.org/10.1007/BF02293814

Avvisati, F., & Keslair, F. (2014). REPEST: Stata module to run estimations with weighted replicate samples and plausible values (Version revised 05 Jun 2019). https://ideas.repec.org/c/boc/bocode/s457918.html

Baker, E. L. & R. E. Mayer (1999). Computer-based assessment of problem solving. *Computers in Human Behavior 15*(3-4): 269-282.

Bear, J. B. & Woolley, A. W. (2011), The role of gender in team collaboration and performance, *Interdisciplinary Science Reviews, 36*, 2-14. https://doi.org/10.1179/030801811X13013181961473

Çakır, M. P., Zemel, A., & Stahl, G. (2009). The joint organization of interaction within a multimodal CSCL medium. *International Journal of Computer-Supported Collaborative Learning, 4*(2), 115–149. https://doi.org/10.1007/s11412-009-9061-0

Care, E., Scoular, C., & Griffin, P. (2016). Assessment of collaborative problem solving in education environments. *Applied Measurement in Education. 29*(4), 250-264. https://doi.org/10.1080/08957347.2016.1209204

Chung, G. K. & Baker, E. L. (2003). An exploratory study to examine the feasibility of measuring problem-solving processes using a click-through interface. *Journal of Technology, Learning, and Assessment 2*(2). https://ejournals.bc.edu/index.php/jtla/article/view/1662

Cloney, D., & Adams, R. J. (2020). *Conquestr version 0.8.3.* https://cran.r-project.org/package=conquestr

Dehler, J., Bodemer, D., & Buder, J. (2007). Fostering audience design of computer-mediated knowledge communication by knowledge mirroring. In C. Chinn, G. Erkens, & S. Puntambekar (Eds.), *Proceedings of the 7th Computer Supported Collaborative Learning Conference* (pp. 168-170). International Society of the Learning Sciences, Inc.

Dillenbourg, P. (1999). What do you mean by ‘collaborative learning’? In P. Dillenbourg (Ed.) Collaborative-learning: Cognitive and computational approaches (pp. 1–19). Elsevier.

Graesser, A. C., Foltz, P. W., Rosen, Y., Shaffer, D. W., Forsyth, C., & Germany, M.-L. (2018). Challenges of Assessing Collaborative Problem Solving. In *Assessment and Teaching of 21st Century Skills* (pp. 75–91). Springer. https://doi.org/10.1007/978-3-319-65368-6_5

Griffin, P., McGaw, B., & Care, E. (Eds.) (2012). *Assessment and Teaching of 21st Century Skills*. Springer. https://doi.org/10.1007/978-94-007-2324-5

Griffin, P., & Care, E. (2015). *Assessment and teaching 21st century skills: Methods and approach.* Springer. https://doi.org/10.1007/978-94-017-9395-7

Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (2020). 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 Approach* (pp. 37-56). Springer. https://doi.org/10.1007/978-94-017-9395-7_2

Hsieh, I.-L. G., & O’Neil, H. F. (2002). Types of feedback in a computer-based collaborative problem-solving group task. *Computers in Human Behavior, 18*(6), 699–715. https://doi.org/10.1016/S0747-5632(02)00025-0

Lin, K.-Y., Yu, K.-C., Hsiao, H.-S., Chu, Y.-H., Chang, Y.-S., & Chien, Y.-H. (2015). Design of an assessment system for collaborative problem solving in STEM education. *Journal of Computers in Education, 2*(3), 301–322. https://doi.org/10.1007/s40692-015-0038-x

McGivney, S., Smeaton, A.F., & Lee, H. (2008). The effect of personality on collaborative task performance and interaction. In E. Bertino& J. B. D. Joshe (Eds.), *Collaborative Computing: Networking, Applications and Worksharing* (pp. 499-511). Springer. https://doi.org/10.1007/978-3-642-03354-4_38

Meier, A., Spada, H., & Rummel, N. (2007). A rating scheme for assessing the quality of computer-supported collaboration process. *Computer-Supported Collaborative Learning, 2*, 63-86. https://doi.org/10.1007/s11412-006-9005-x

Mendelovits, J. (2017). Test Development. In P. Lietz, J. C. Cresswell, K. F. Rust & R. J. Adams (Eds) Implementation of Large-Scale Education Assessments (pp. 63-91). John Wiley and Sons Ltd. https://doi.org/10.1002/9781118762462.ch3

O’Neil, H. F., Chuang, S., & Chung, G. K. W. K. (2004). Issues in the computer-based assessment of collaborative problem solving. *Assessment in Education, 10*, 361-373. https://doi.org/10.1080/0969594032000148190

Okada, T., & Simon, H. A. (1997). Collaborative discovery in a scientific domain. *Cognitive Science, 21*(2), 107-146. https://doi.org/10.1207/s15516709cog2102_1

OECD (2017), "PISA 2015 collaborative problem-solving framework", In *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving* (pp.131-188). OECD Publishing. https://doi.org/10.1787/9789264281820-8-en.

Organisation for Economic Co-operation and Development (OECD). (2017a). *PISA 2015 Technical Report*. OECD Publishing. http://www.oecd.org/pisa/data/2015-technical-report/

Organisation for Economic Co-operation and Development (OECD). (2017b). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. OECD Publishing. https://doi.org/10.1787/9789264281820-en

Organisation for Economic Co-operation and Development (OECD). (2017c). *PISA 2015 Results (Volume V): Collaborative problem solving*. OECD Publishing. https://doi.org/10.1787/9789264285521-en

Palincsar, A. S., & Magnusson, S. J. (2001). The interplay of first-hand and text-based investigations to model and support the development of scientific knowledge and reasoning. In S. Carver & D. Klahr (Eds.)*, Cognition and instruction: Twenty-five years of progress* (pp. 151-193). Lawrence Erlbaum Associates.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* National Research Council. National Academy Press.

R Core Team. (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Ramalingam, D. & Adams, R. J. (2017). How Can the Use of Data from Computer-Delivered Assessments Improve the Measurement of Twenty-First Century Skills. *Assessment and Teaching of 21st Century Skills: Research and Applications*, E. Care, P. Griffin and M. Wilson (Eds.) (pp. 225-238). Springer. https://doi.org/10.1007/978-3-319-65368-6_13

Rasch, G. (1960). *Probalistic Models for Some Intelligence and Attainment Tests*. Danmarks Pædagogiske Institut.

Schacter, J., Herl, H. E., Chung, G. K. W. K., Dennis, R. A., & O'Neil, H.F. (1999). Computer-Based Performance Assessments: A Solution to the Narrow Measurement and Reporting of Problem-Solving. *Computers in Human Behavior 15*(3-4): 403-418.

Scoular, C., Care, E., & Hesse, F. (2017). Designs for operationalizing collaborative problem solving for automated assessment. *Journal of Educational Measurement. 54*(1), 12-35. https://doi.org/10.1111/jedm.12130

Scoular, C. (2019). A design template for transforming games into twenty-first century skills assessments. *Journal of Applied Research in Higher Education.* https://doi.org/10.1108/JARHE-02-2018-0018

Scoular, C., & Care, E. (2019). Monitoring patterns of social and cognitive student behaviors in online collaborative problem solving assessments. *Computers in Human Behavior, 105874.* https://doi.org/10.1016/j.chb.2019.01.007

Scoular, C., Duckworth, D., Heard, J., & Ramalingam, D. (2020). *Collaboration: Definition and structure.* Australian Council for Educational Research. https://research.acer.edu.au/ar_misc/39

Scoular, C. & Timms, M. J. (2020). Development of a Measurement Approach to Assess 21st Century Skills. *Contemporary Perspectives on Research in Educational Assessment*, (pp.55-66). Information Age Publishing.

Singh-Gupta, V., & Troutt-Ervin, E. (1996). Preparing students for teamwork through collaborative writing and peer review techniques. *Teaching English in the Two Year College, 23,* 127-136.

Teasley, S. D. (1995). Communication and collaboration: The role of talk in children's peer interactions. *Developmental Psychology, 31*(2), 207-220.

Von Davier, A. A., & Halpin, P. F. (2013). *Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations.* Research Reports: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02348.x

Wildman, J.L., Shuffler, M. L., Lazzara, E., & Garven, S. (2012). Trust development in swift starting action teams: A multilevel framework. *Group & Organization Management*, *37*, 138-170. https://doi.org/10.1177/1059601111434202

Wright, B., & Masters, G. (1982). *Rating scale analysis.* IL: MESA Press. https://research.acer.edu.au/measurement/2

Wu, M. L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalised item response models.* http://hdl.handle.net/11343/57769

Zagal, J. P., Rick, J., & His, I. (2006). Collaborative games: Lessons learned from board games. *Simulation and Gaming, 37*(1), 24-40. https://doi.org/10.1177/1046878105282279

Zoanetti, N. (2010). Interactive computer based assessment tasks: How problem-solving process data can inform instruction. *Australasian Journal of Educational Technology*, *26*(5), 585-606. https://doi.org/10.14742/ajet.1053

Acknowledgements

The contribution of Sofia Eleftheriadou was supported by the Economic and Social Research Council (ESRC) (NWSSDTP grant number ES/P000665/1) during her residency at the Australian Council for Educational Research.

Appendix

INSERT APPENDIX TABLE HERE

Table 1. PISA 2015 CPS framework (OECD, 2017b)

	(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 the 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) Describing roles and team organisation (communication protocol/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

Table 2. Features of each assessment

Feature	PISA-CPS	DHO-CPS	ACER-GC
Purpose	International comparison; population-level statistics	Formative assessment	Formative assessment
Number of Australian students	4305 students	1080 students	1145 students
Target population	15-year old's	11-16-year old's	10- & 15-year old's
Length	1 hour	1.5 hours	4 hours
Mode	Human-Computer Agent	Human-Human (2)	Human-Human (3)
Action types	Selecting options, dragging and dropping resources, chat box communication	Selecting options, dragging and dropping resources, sharing resources, chat box communication	Selecting options, entering text, chat box communication
Target measure	Individual	Individual	Individual and group
Data capture	Closed response selection	Process data, chat logs, closed and open response selection	Chat logs, process data, open response selection
Resource symmetry	Asymmetrical	Symmetrical	Asymmetrical

Table 3. Distribution of the construct across assessments

Aspects of ACER collaboration framework	**PISA-CPS**	**DHO-CPS**	**ACER-GC**
1.1 Communicates with others	2	18	3
1.2 Pools resources and information	-	6	1
1.3 Negotiates roles and responsibilities	-	-	1
2.1 Participates in the group	1	6	1
2.2 Recognises contributions of others	2	6	2
2.3 Engages with role and responsibilities	3	6	-
3.1 Ensures constructiveness of own contributions	2	9	3
3.2 Resolves differences	1	6	4
3.3 Maintains shared understanding	1	15	1
3.4 Adapts behaviour and contributions for others	-	-	-
Total	12[1]	72	16

[1] Only 12 items out of 117 from the PISA-CPS were included in the mapping as the content was required and these are the only publicly available items

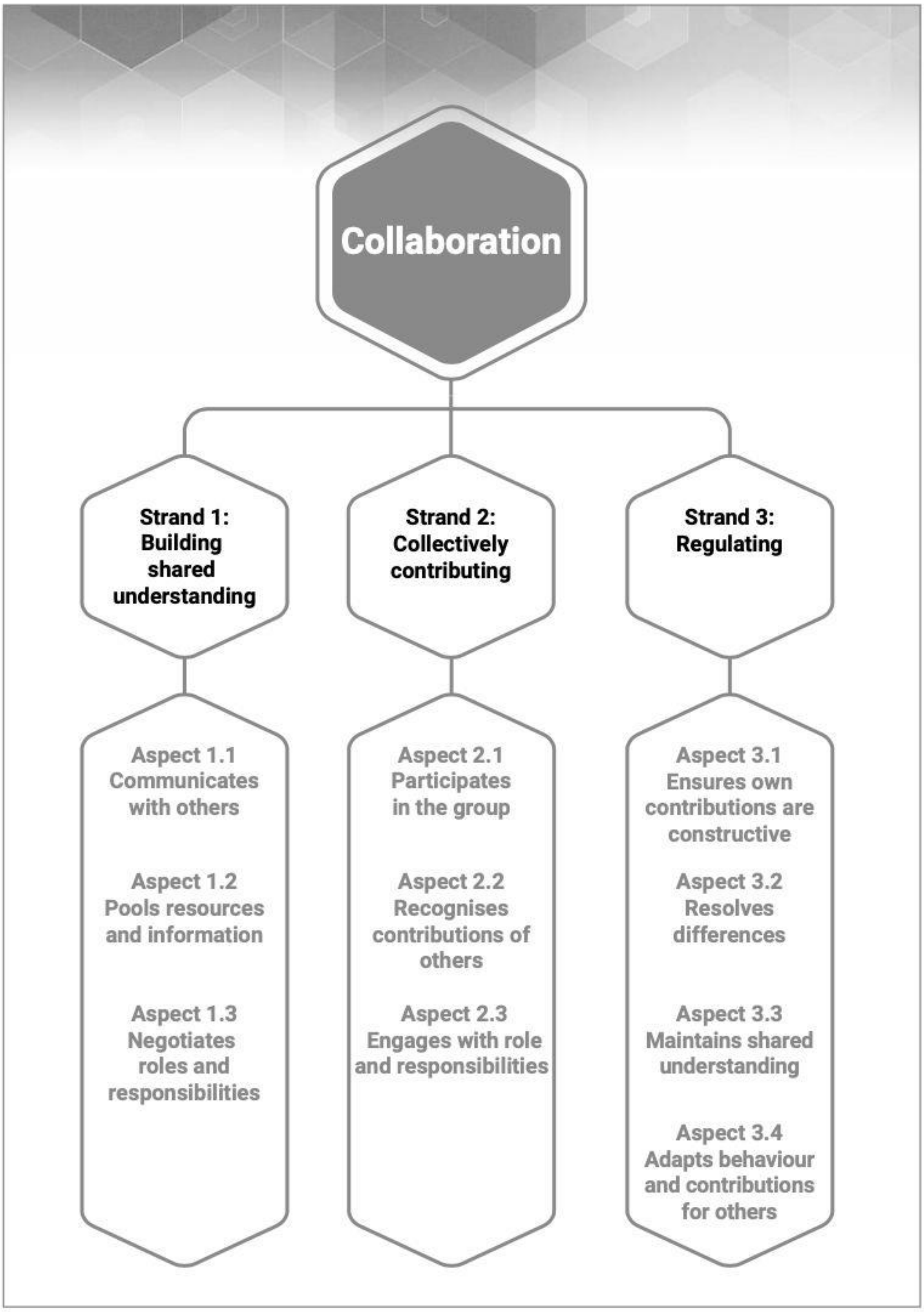
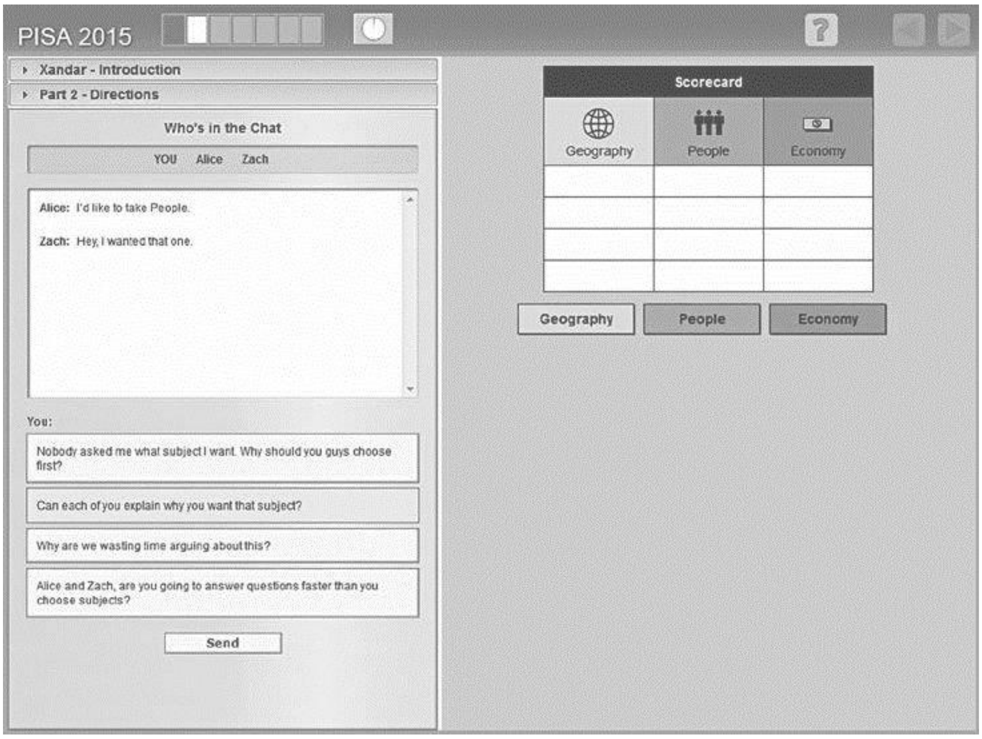

Collaboration
Strand 1: Building shared understanding
Strand 2: Collectively contributing
Strand 3: Regulating
Aspect 1.1 Communicates with others
Aspect 1.2 Pools resources and information
Aspect 1.3 Negotiates roles and responsibilities
Aspect 2.1 Participates in the group
Aspect 2.2 Recognises contributions of others
Aspect 2.3 Engages with role and responsibilities
Aspect 3.1 Ensures own contributions are constructive
Aspect 3.2 Resolves differences
Aspect 3.3 Maintains shared understanding
Aspect 3.4 Adapts behaviour and contributions for others

Figure 1. ACER Framework for Collaboration (Scoular et al., 2020)



Alice: I'd like to take People.

Zach: Hey, I wanted that one.

You:

Nobody asked me what subject I want. Why should you guys choose first?

Can each of you explain why you want that subject?

Why are we wasting time arguing about this?

Alice and Zach, are you going to answer questions faster than you choose subjects?

Figure 2. Screenshot of a released PISA-CPS item (from OECD, 2017c)

Note: Chat space (left) displays the pre-defined messages for communication with the computer agents, and task space (right) where actions are performed. Second message (highlighted) is the credited response

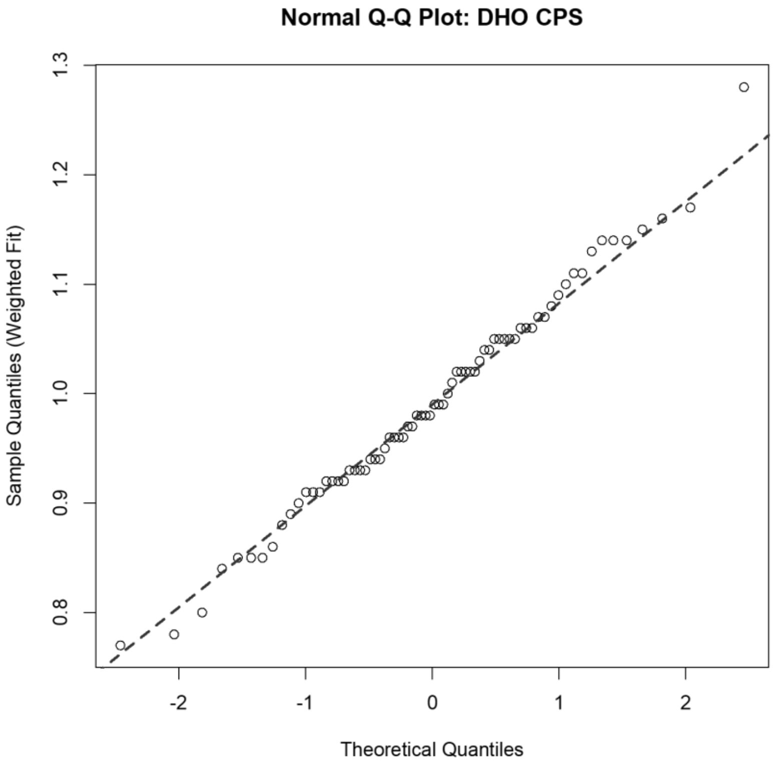


Figure 3. Normal quantile-quantile (QQ) plot of DHO-CPS weighted fit statistics (dashed line shows line passing through 25th and 75th centiles)

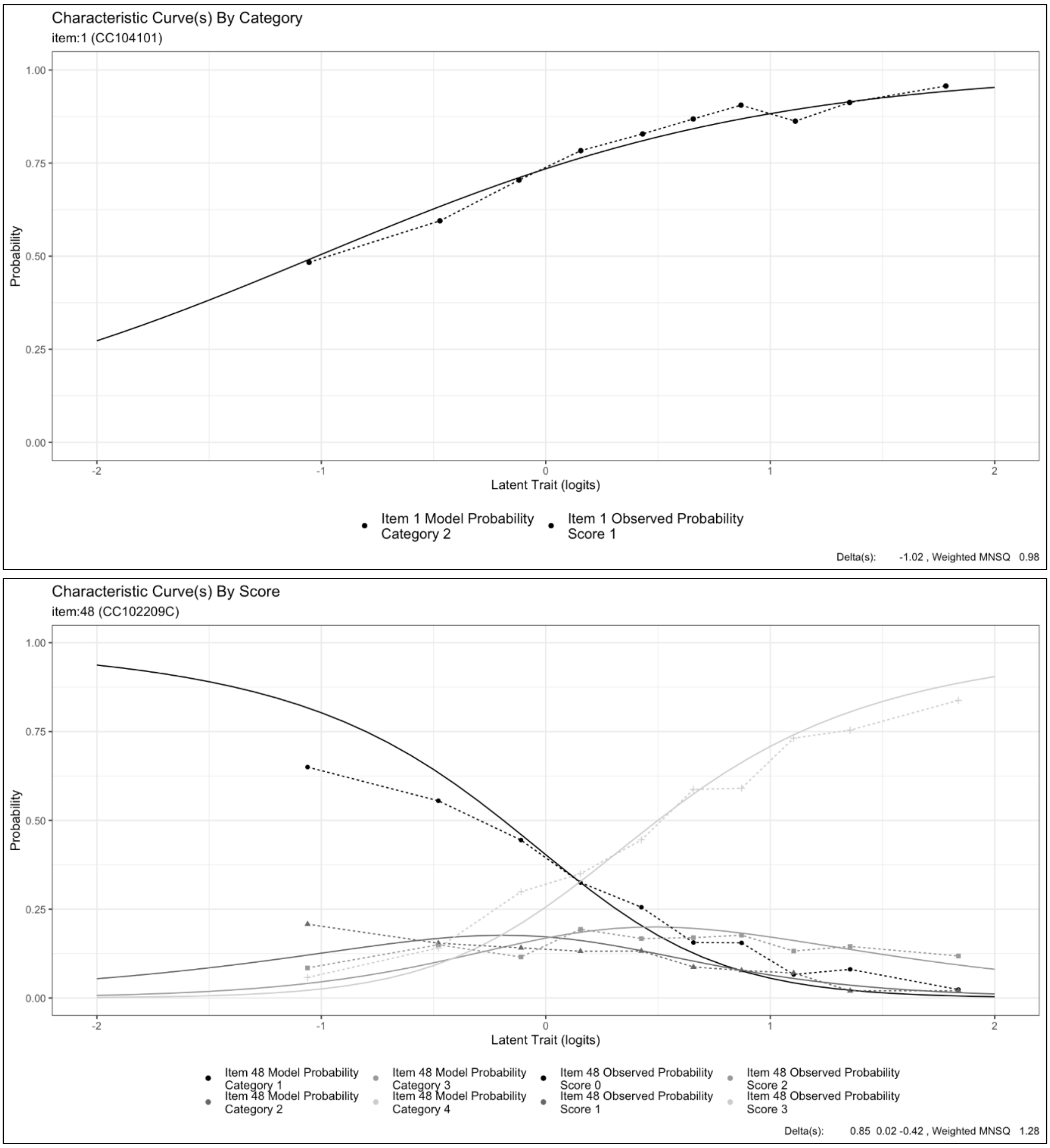


Figure 4. Item characteristic curves showing a good fitting item (top) and a poorer fitting item (bottom) from the PISA-CPS scale

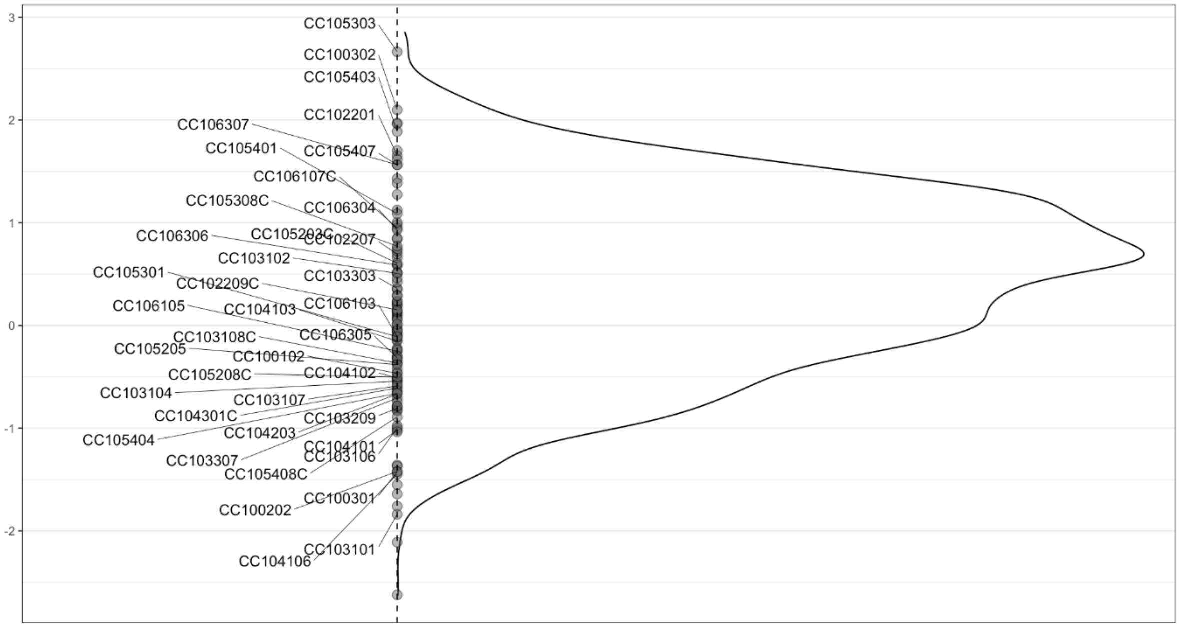


Figure 5. Wright Map of item locations on the PISA-CPS scale.

Note: Some item labels have been omitted to minimise overlapping