*AERA Open*
*July-September 2017, Vol. 3, No. 3, pp. 1–19*
*DOI: 10.1177/2332858417728046*


# Measuring Collaborative Problem Solving Using Mathematics-Based Tasks

**Susan-Marie E. Harding**
**Patrick E. Griffin**
**Nafisa Awwal**
**BM M. Alom**
**Claire Scoular**
*University of Melbourne*

*This study describes an online method of measuring individual students' collaborative problem-solving abilities using four interactive mathematics-based tasks, with students working in pairs. Process stream data were captured from 3,000 students who completed the tasks in the United States, Australia, Canada, Costa Rica, Singapore, and Finland. The data were transformed into indicators of collaborative problem-solving ability and were analyzed using item response modeling. The assessments employed in this study can be used as a teaching tool for introduction to algebraic concepts and as a measurement instrument for collaborative problem-solving ability. The paper describes the construction, calibration, and reliability of the tasks and considers validation issues, such as fairness between assessments for both partners and avoidance of cultural biases. Investigations into the dependencies between student scores provide evidence for convergent and discriminant validity.*

Keywords: *collaboration, problem solving, mathematics, secondary education, assessment, 21st century skills, collaborative problem solving*

Collaborative problem solving (CPS) has been described as a critical skill for students to develop, but the nature of communication between individuals in a collaborative situation causes difficulty in measuring students' ability using traditional testing approaches, such as multiple choice, extended answer, peer review, or teacher observation. Further, when students are scored on a group task, teachers often allocate one score to the whole group of students, disregarding individual students' performances, and that score might reflect achievement of the correct solution rather than contribution to the collaborative process. Another issue is that teachers who wish to assess CPS in their classrooms may not have the scope to do so if this requires deviating from teaching the mandated curriculum. Such issues in assessing CPS arise regardless of the teachers' specific content area specialization.

This paper considers collaborations in mathematics classrooms in junior high school or middle school (with students in approximately Years [or Grades] 6, 7, and 8) in six different countries. Students in this age range commonly study basic algebra and the concepts of variables, integers, and polynomials. Their teachers need to manage the transition from arithmetic to algebra and accommodate current curriculum reforms emphasizing capabilities such as CPS and critical thinking. These latter capabilities require accurate and meaningful assessment, just as the more familiar content areas do.

These issues and complications in measuring CPS while teaching a content domain underlie the overarching research question addressed in this study: How can teachers assess CPS for students individually while they teach mathematics?

## Why Collaborative Problem Solving?

Problem solving has been assessed for several decades, following Polya's (1945/1973) publication of his four-step process within the domain of mathematics. As this process was embedded within mathematics curricula, problem solving was generally regarded as a mathematics skill. Studies in the latter part of the 20th century and in the early part of the 21st century linked problem solving with decision making, critical thinking, and collaboration (e.g., O'Neil, 1999; O'Neil, Chuang, & Chung, 2003; Griffin, Care, & McGaw, 2012). The inclusion of these skills extended the exploration of problem solving beyond the field of mathematics into the sciences and other discipline areas, and the skills are now broadly considered to be independent of those disciplines and a subject of learning in their own right. The combination

of these skills as CPS was formalized in the project titled the Assessment and Teaching of 21st Century Skills (ATC21S) (Griffin et al., 2012; Griffin & Care, 2015).

Increasing awareness of the importance of team participation in the workplace (e.g., Economist Intelligence Unit, 2015) and the linking of problem solving and critical thinking to increased productivity (World Economic Forum, 2016) have led to a realization that CPS has a central role in learning and work in the 21st century. Because of the shift from an industrial economy to a knowledge- and information-based economy, there is an increasing demand for workers who have developed strong collaboration, critical-thinking, information technology, and problem-solving skills during their education (Greiff, Holt, & Funke, 2013; Griffin et al., 2012; O'Neil et al., 2003). Such is the recognition given to CPS as a valued life skill that the Organisation for Economic Co-operation and Development (OECD) conducted a comprehensive study of it in 53 countries as part of the Programme for International Student Assessment (PISA) in 2015. Many research groups around the globe are in the process of creating tasks and investigating data analytic techniques for CPS assessments (von Davier, Zhu, & Kyllonen, 2017). The focus appears to be on how to best assess the construct rather than on how teachers could best use CPS assessments in class.

At a policy level, education is shifting toward an emphasis on generic competencies, and the involvement of the OECD is expected to accelerate pressure for curricula to include CPS and other 21st century skills. To embed these kinds of skills into curricula, new teaching and assessment practices are needed. Data from assessments are needed in order to inform teaching and to indicate the kinds of skills that students need to learn in order to become more proficient in collaborative practices. This is now widely accepted as an important preparation for participation in the workforce, and many countries have begun adapting their curriculum to incorporate 21st century skills (referred to variously as "transdisciplinary" skills or competencies, "general capabilities," "work skills," "soft skills," and so on) into the traditional key learning areas (Care, Anderson, & Kim, 2016). This has encouraged the development of assessment strategies to assist teachers to deal with the curriculum shift.

However, teachers have identified practical difficulties in incorporating 21st century skills into classroom teaching. According to global survey results published by the Economist Intelligence Unit (2015), 49% of teachers report that lack of time in a strictly regulated curriculum is the biggest barrier to teaching 21st century skills. Another 30% list education authorities' strict requirements that focus in the classroom be on literacy and numeracy as the biggest challenge faced. This paper seeks to address these problems by demonstrating that the use of CPS assessments housed within mathematics algebra tasks allows teachers to use their time to teach the curriculum while also assessing CPS skills. And this development has wider implications: Other 21st century skills could be taught in a similar way, alleviating more of the pressures that teachers feel in introducing the assessment of these types of skills into their classrooms.

### *Theoretical Framework*

The ATC21S project pioneered the development of an online human-to-human CPS assessment and established a benchmark for the construction of tasks and the interpretation of student performances. A defining feature of the ATC21S tasks was the necessity to allocate separate resources to each of the students, so that no student could solve the problem alone and no student could be a nonparticipant in the process. The problem could be resolved only if each student contributed the resources he or she controlled and actively participated in the process. CPS was defined as the process of approaching a problem responsively by working together and exchanging ideas, but Griffin (2014) has more recently refined the definition as

> a joint activity where two or more people work together to contribute resources they alone control, to progress through a series of cognitive states that involve collection and analysis of information and the formulation of hypotheses that they jointly set out to test. (p. 12)

CPS differs from individual problem solving in that much of the students' activity is overt. Griffin (2014) described the difference between individual problem solving and CPS in the following way:

> The primary distinction between problem-solving by an individual and collaborative problem-solving is its social nature—the need for communication, exchange of ideas, shared identification of the problem and its elements, and negotiated agreement on connections between problem elements and relationships between actions and their effects. Collaborative problem-solving makes each of these steps observable, as they must be shared with a partner or other members of a group if a solution is to be successfully identified. (p. 9)

This article describes an analysis of student performances on mathematics-related CPS tasks that were created as part of the ATC21S project. The tasks were created in English and completed in that language by students in Australia, the United States of America, and Singapore. The tasks were translated into Spanish for students in Costa Rica, Finnish for students in Finland, and Dutch for students in the Netherlands.

The complex construct of CPS used to create the tasks was described by Hesse and others (Hesse, Care, Buder, Sassenberg, & Griffin, 2015). They proposed a CPS framework consisting of social and cognitive strands and the definition of the variable. The indicative behaviors measured in the CPS assessments were classified as belonging to either social or cognitive skill areas. Social skills were further

TABLE 1
*Collaborative Problem-Solving Skills and Indicator Design*

| Element | Behavioural Indicators Based on . . . |
|---|---|
| Social skills | |
| Participation | |
| Action | Activity within environment |
| Interaction | Interacting with, prompting, and responding to contributions of others |
| Task completion/perseverance | Undertaking and completing a task or part of a task individually |
| Perspective taking | |
| Adaptive responsiveness | Ignoring, accepting, or adapting contributions of others |
| Audience awareness (mutual modeling) | Awareness of how to adapt behavior to increase suitability for others |
| Social regulation | |
| Negotiation | Achieving a resolution or reaching compromise |
| Self-evaluation (meta-memory) | Recognizing own strengths and weaknesses |
| Transactive memory | Recognizing the strengths and weaknesses of others |
| Responsibility initiative | Assuming responsibility for ensuring parts of the task are completed by the group |
| Cognitive skills | |
| Task regulation | |
| Organizes (problem analysis) | Analyzing and describing a problem in familiar language |
| Sets goals | Setting a clear goal for the task |
| Resource management | Managing resources or people to complete a task |
| Flexibility and ambiguity | Accepting ambiguous situations |
| Collects information | Exploring and understanding elements of the task |
| Systematicity | Trying possible solutions to a problem and monitoring progress |
| Learning and knowledge building | |
| Relationships (represents and formulates) | Identifying connections and patterns between and among elements of knowledge |
| Contingencies/rules (“If . . . then”) | Using understanding of cause and effect to develop a plan |
| Hypothesis (“What if . . .”) (reflects and monitors) | Adapting reasoning or course of action as information or circumstances change |

*Source*. Hesse, Care, Buder, Sassenberg, and Griffin (2015).

classified as *participation* (action, interaction, persistence), *perspective taking* (adaptive responsiveness, audience awareness), or *social regulation* (negotiation, meta-memory, transacted memory, responsibility initiative), while cognitive skills were classified as *task regulation* (problem analysis, goal setting, resource management or control, flexibility in an ambiguous context, data collection, systematicity) or *learning and knowledge building* (relationships, contingencies and generalization, hypothesis testing). To explore the log stream data for evidence that students were displaying these skills, indicators were designed to represent the elements described in Table 1. The tasks were completed by student dyads, and the assessment was designed to record the process the students used to solve the problems, including the social and cognitive skills they used, in addition to a problem solution. This emphasis on the process used by students was adopted in anticipation of the need for teachers to understand the process involved when students learn to develop the skills of CPS and to be able to translate that understanding into pedagogical practice.

The mathematics-based tasks used in this study introduce students to algebraic concepts using numerical reasoning (rich tasks) in an interactive way designed to be engaging. The tasks can be used to teach mathematics and meet curriculum demands for both mathematics content and 21st century skills. In addition to emphasising the problem-solving process, the tasks of the ATC21S project allow assessment of individual student performance within the composition of a team or group collaboratively solving problems.

### *An Individual Measure for CPS*

In the editorial of a special issue on collaborative educational assessments in the *Journal of Education Measurement*, von Davier (2017) argued that “in modeling the *process* of collaboration, we are concerned about describing the statistical dependence exhibited by the activities of groups of individuals” (p. 8). This approach, described by von Davier and Halpin (2013) and others,

models the statistical dependencies between individuals and considers the degree of dependency as a measure of collaboration. In contrast, the process described in this paper attempts to control for statistical dependency between partners by measuring only the individual's contribution to the collaborative process. The method of scoring each individual on separate indicators, other than those shown to be statistically common between partners, limits dependencies to provide an individual measure of CPS ability.

In its draft framework for individual problem solving, PISA (2010) identified three challenges for the assessment of CPS skills: how to assign credit to individual group members if this is required, how to account for differences across groups that may bias individual performance, and how to account for cultural differences in group dynamics (PISA, 2010). The first challenge is addressed in this study. However, the separate questions of how to account for differences in group dynamics and cultural differences are not addressed, nor is the theoretical rationale behind wanting or needing to account for any real differences in ability between cultures or groups.

Previous publications on the creation, coding, scoring, calibration, and use of the ATC21S tasks have provided various forms of evidence for the validity and reliability of the assessments. A set of four tasks was described in detail, with an outline of how each of the tasks covers the skills included in the Hesse framework, presenting evidence for validation in terms of instrument content (Care, Griffin, Scoular, Awwal, & Zoanetti, 2015). The process of the coding of the data into indicators was described, demonstrating validation for response processess (Adams et al., 2015). The characteristics of the full set of 11 tasks have been described, including concept and construct mapping, specifications, blueprints, and task calibration, establishing the internal structure of the assessments (Griffin, Care, & Harding, 2015). Evidence for the reliability and validity of a different set of four CPS tasks (a mixture of content-free and content-dependent tasks), including precision data and test information curves, has been reported in Harding and Griffin (2016), where analyses of the data demonstrated that the performance of the tasks and the indicators of student behavior were independent of language, curriculum, or national emphasis. The coding and scoring mechanisms presented in this paper are the same as described in these articles.

### *Research Questions*

The purpose of this paper is to extend the understanding of the possible uses of CPS assessments by exploring the benefits and limitations of housing CPS tasks in the specific curriculum content of mathematics. Of particular importance, here we describe the exact mathematical skills covered by the four tasks to demonstrate how they may be useful for teaching specific mathematical concepts. In designing our research questions, we aimed to address the validity, reliability, and fairness of the tasks with reference to the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014). Analyses of the data were conducted to ensure that the set of mathematics CPS tasks adequately covered the construct, ensuring that they are reliable as a measure of CPS. Consequential validation was considered by examining fairness between different test forms (Students A and B) and cultural bias. Investigations of the dependency of one partner's social or cognitive CPS skills on the other were conducted to provide evidence of discriminant validity. Scores between partners were not expected to correlate, as the scoring was designed to provide an individual measure of CPS ability. To examine discriminant and convergent validity, students' CPS performance on the mathematical tasks was compared to their performance on a set of content-free tasks. We hypothesed that student estimates on the social skills for the mathemematics tasks would correlate with the estimates on the social skills for the content-free tasks, providing evidence for convergent validity. The students' ability estimates on the cognitive dimension were hypothesized to differ depending on the content area. However, the estimates were expected to correlate to some degree. The aim of these analyses was to begin investigation of the potential dependencies that need to be considered by teachers when evaluating a student's level of performance on CPS tasks and to raise considerations for improvements in the creation and analytical techniques of assessments of CPS in the future.

The research questions addressed in this paper are as follows:

1. Can an assessment for human-to-human CPS be framed within a mathematical domain and provide accurate estimates of ability for students, while being of use in teaching mathematics?
2. Is the set of mathematics CPS tasks fair for each student A and B, and do the tasks avoid cultural bias?
3. What is the dependency of one partner's social and cognitive CPS skills on the other?
4. What is the relationship between the students' ability estimates in the mathematics and the content-free CPS tasks?

## Method

### *Sample*

Data were obtained from 3,004 students (ages 11–17) from Australia ($n = 554$), Costa Rica ($n = 362$), the Netherlands ($n = 288$), Finland ($n = 306$), Singapore ($n = 878$), and the United States ($n = 556$). Students completed a selection from a suite of tasks, including Hexagons ($n = 846$), Warehouse

($n$ = 1,922), Game of 20 ($n$ = 1,184), and Small Pyramids ($n$ = 995). The samples of students from each country were opportunistic and not representative of the total population. The country mean scores could therefore not be compared legitimately as indicators of the skill level of students in that country. However, the psychometric properties of indicators across countries were compared by ordering and determining the correlation of the difficulty parameters of items. The student pairings were not randomized. Each student pair was allocated by class within each country for test administration purposes. There is a probable bias toward matched students of similar ability because the students in each pairing came from the same class, within the same (approximate) year or grade level, and within the same country. As there was no systematic application of a randomized sampling method, any findings based on student pairings have potential limitations, which will be discussed.

### *The Tasks*

The mathematics tasks used are titled Hexagons, Warehouse, Game of 20, and Small Pyramids. To address the fourth research question involving comparison with non-mathematical tasks, the content-free tasks used were Olive Oil, Clown, Shared Garden, Plant Growth, and Sunflower (described by Care et al., 2015). The content-free tasks required no prior content knowledge on any particular subject. All the tasks involve two students (Student A and Student B) working through problems together, on separate computers, and communicating information via a chat box that functions in a similar way to text messaging. The mechanism for communication is solely via the chat box, allowing all collaboration to be explicit and recorded. Each student is given a unique set of resources to use in working through the problem, so students need to ask their partner for information to fully understand the problem space and to gather the information required.

Each task includes an introductory screen, followed by several "pages" of subtasks or problems to complete. The introductory screen of each task follows a similar format: A brief description of the task is included along with a basic description of what is required of the student in completing the task. After completion of the tasks, the students were also required to answer survey questions, where they were required to judge their own and their partner's performance. Peer learning may occur in the tasks if a strong student is paired with a weaker student. After the completion of this task, the teacher may take the opportunity to create a discussion about how the task could have been solved and how the rule may have been produced.

*Hexagons.* In Hexagons, students work together to find the arithmetical patterns that govern the workings of a puzzle in which numbers appear in small hexagons within a larger triangle. The task has eight problem-solving screens that require no algebraic skills; it involves multiplication and addition patterns with negative and positive numbers. Students manipulate the number pattern by choosing the starting number and/or the value change for the number pattern (e.g., +4). The task shares some similarities with the more common input-output "machines" that some mathematics teachers use to introduce or allow practice in solving number patterns and determining the rules that govern them. The first page of the task allows students to alter the numbers in the triangle and to investigate how the number pattern works. In this task, Student A controls the rule down the left side of the triangle, and Student B controls the rule down the right side, as illustrated in Figure 1. On the right side of the screen, a chat box allows students to talk to each other to discuss the problem. Students are guided throughout the different pages in the task to create their own rule for the triangle and to discover the rule their partner has created. They are guided with the visual aid of the triangle for the first six pages, after which the students are asked to solve the number patterns without the aid of the diagram, as shown in Figure 2. Hexagons is most suitable for students from Year/Grade 6 (age 11) to Year/Grade 8 (age 13).

*Warehouse.* In this task, students are asked to secure a warehouse by correctly positioning security cameras around tall boxes that block the cameras' view. Students need to assess how the cameras behave and find the rule that determines the minimum number of cameras required to secure the warehouse. The five-page task begins with Student A placing cameras around six boxes arranged in a 2-by-3 grid, as illustrated in Figure 3. Student A has control of the cameras but cannot see the areas the cameras cover; Student B can see the areas covered by the cameras but cannot see the cameras themselves. The students need to work together to find out how the cameras operate. The number of boxes in the warehouse increases in subsequent pages, from six (2 by 3) to nine (3 by 3) and then 12 (3 by 4), so students can learn how the rule can be extended to various grid sizes. In the final page of the task, students need to be able to extrapolate the rule they have learned and apply it to larger grids (the largest being 18 by 11) without a visual aid. This task requires strong mathematical reasoning skills. Stronger students will develop an algebraic rule to calculate the number of cameras required, and they will see this as the fastest method. Students who struggle with algebra may attempt to draw the grid to solve the problem. This task can serve as an introduction to algebra for students in Year/Grade 7 or Year/Grade 8.

*Game of 20.* This task has previously been described by Care et al. (2015). It involves two students working as a team and competing with a computer to be the first to arrive at a total of 20 by placing "counters" on a game board with spaces numbered 1, 2, 3, 4, or 5, as shown in Figure 4. When

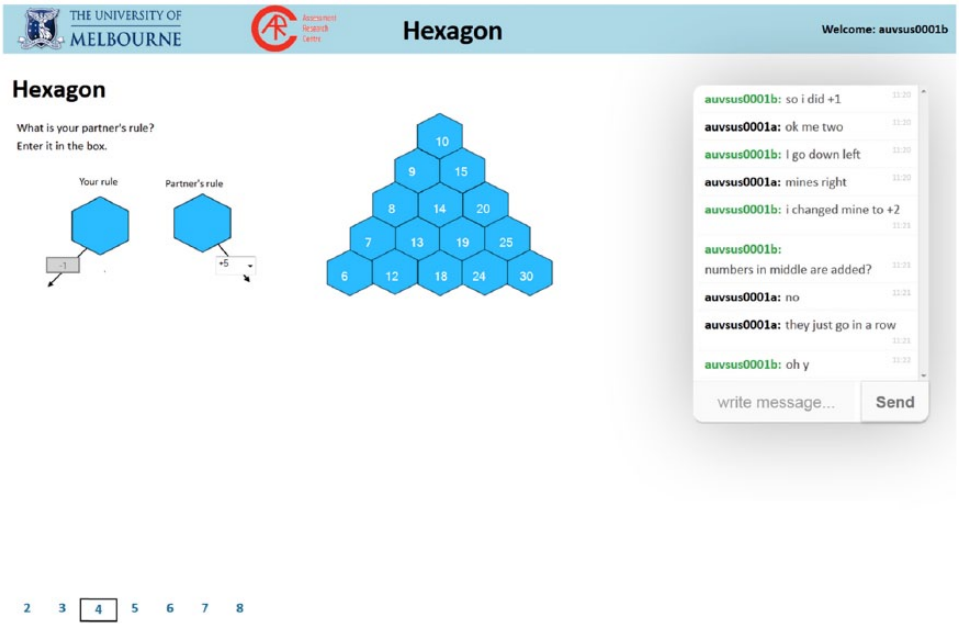


FIGURE 1. *Hexagons task, Student B view, page 4.*

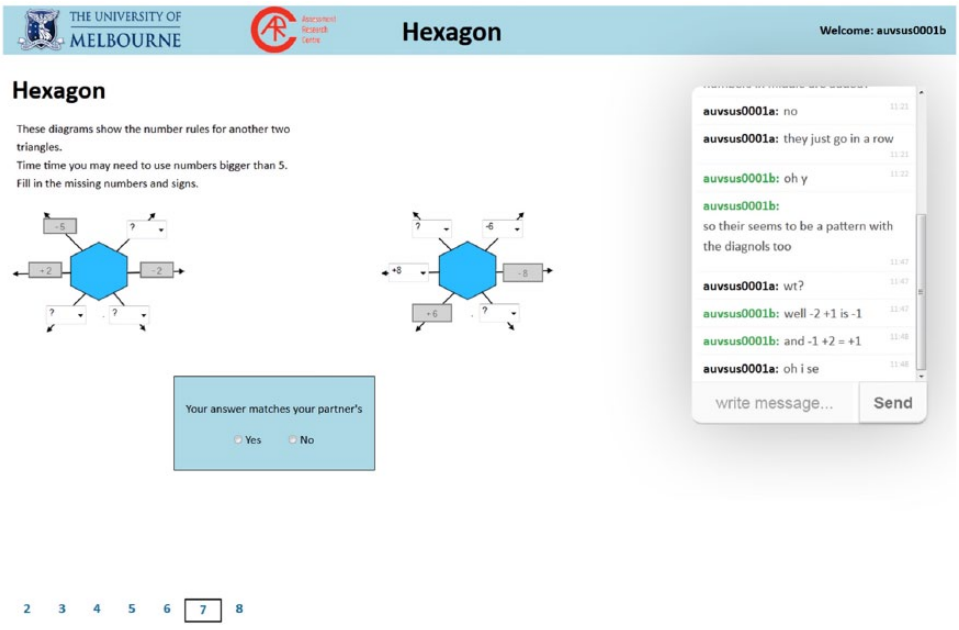


FIGURE 2. *Hexagons task, Student B view, page 7.*

the students have placed a counter on one of the numbers, the computer places a counter on a remaining number. This continues until the total of the covered numbers reaches 20 or the players "bust" (use a counter that takes the total over 20), in which case they lose. The students work together to decide in which number space to place their counter. Before placing the counter, they must each select one number. Student A must select a number between 0 and 4, and Student B must select a number between 1 and 5. The combined total of their two numbers must be no greater than 5: If Student A chooses 3, then Student B can choose only 1 or 2 to make a total of 4 or 5, respectively. When the students have chosen two numbers, the game places a counter on one of the spaces featuring the number of their combined total. As the students play and practice, they are guided to find the totals that will ensure they win the game. They are then asked to specify

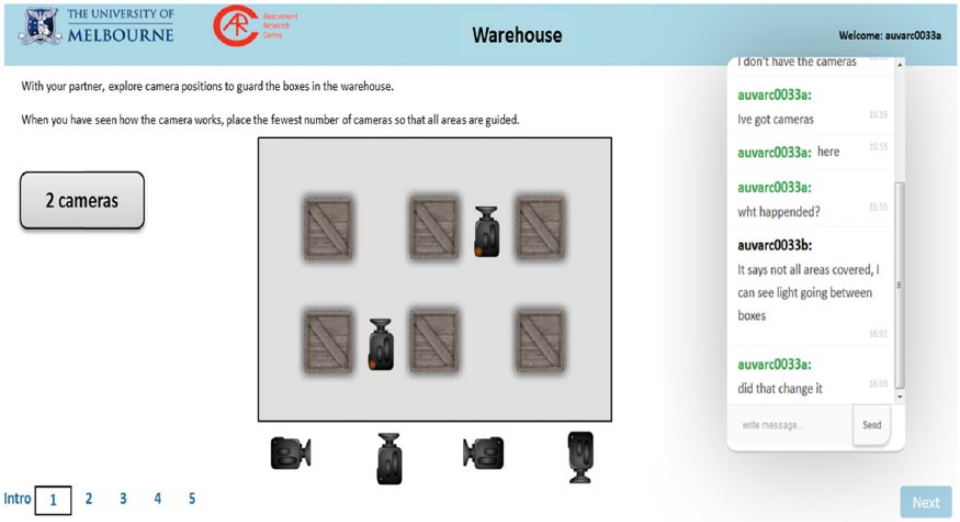


FIGURE 3. *Warehouse task.*

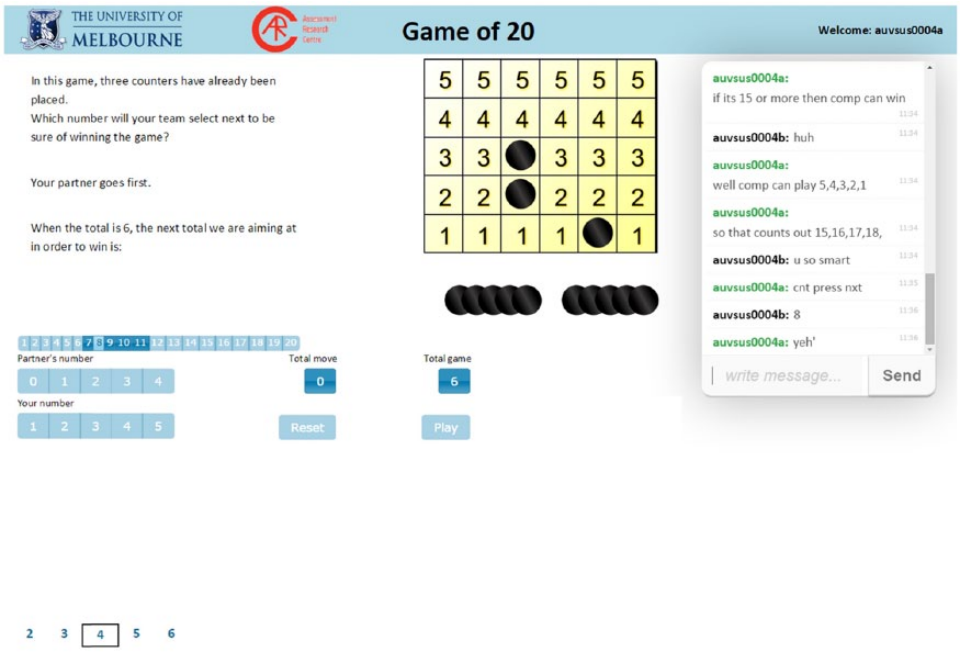


FIGURE 4. *Game of 20 task, Student A view, page 4.*

their winning strategy. The game requires strong collaborative skills, and its application of general numerical reasoning allows it to be used in mathematics classrooms in Years/Grades 6 to 8.

*Small Pyramids.* In this task, students use basic number reasoning skills to determine the rules that apply to the Fibonacci series. The small pyramid is "connected" between Student A's and Student B's screens. When Student A types a number in the red box (bottom left side), Student B can see consequences of the number choice, including the number presented in the black box at the top of the pyramid, about which the student must answer questions, as illustrated in Figure 5. The task is divided into seven pages. The number of boxes filled in the pyramid depends on the level of complexity of the problem presented. On the first page, students explore the pyramid. Student A enters a number in the red box and can see the bottom row of numbers created from that number choice; Student B can see the top half of the pyramid. The students need to communicate the numbers that they can see to each other to find the rule that explains how the boxes are filled. As students continue the task, they

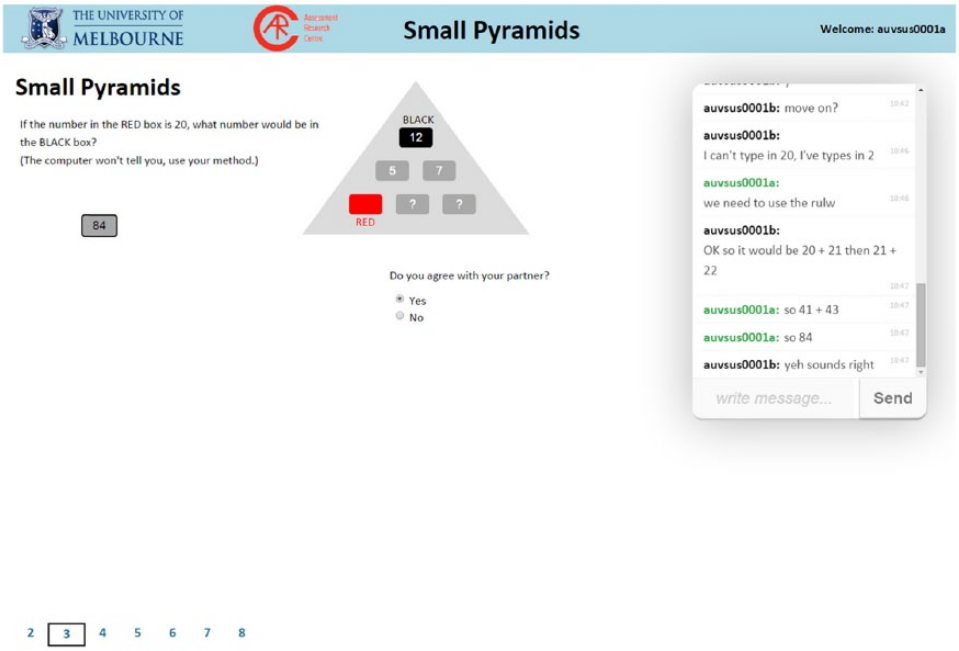


FIGURE 5. *Small Pyramids task, Student B view, page 3.*

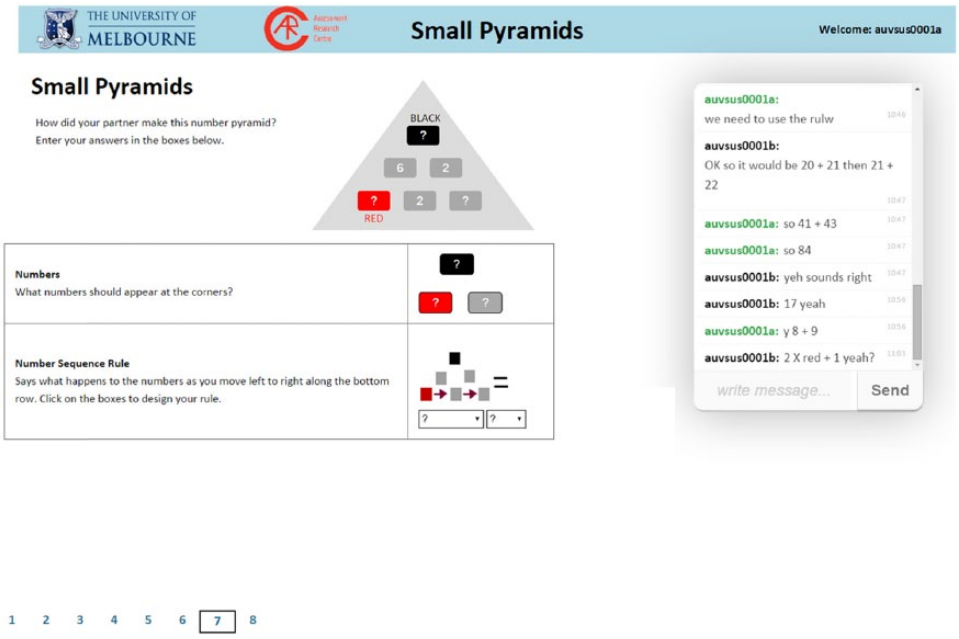


FIGURE 6. *Small Pyramids task, Student B view, page 7.*

are required to predict the numbers that would appear in the boxes of the pyramid, particularly the black box, when a hypothetical number is typed into the red box, as shown in Figure 5. Midway through the pages of the task, students can no longer enter the number from the problem into the red box but need to apply their understanding of the way the pyramid works to solve the presented problem. Eventually, each student is asked to change the rules of the pyramid, and his or her partner is asked to predict the new rules, as shown in Figure 6. When the problem is presented, the students can view only some of the numbers in the pyramid, so by the final page of the task, the students need to have developed a thorough understanding of the relationships between these numbers to identify their partner's rule. This task is suited to

Year/Grade 7 or 8 students as an introduction to algebra or to the use of numerical patterns and mathematical reasoning.

### *Coding and Calibration*

As students completed the tasks, all actions and chats were captured as process stream data and collected in log files, which were then recoded into behavioral indicators designed to represent the elements described in Table 1 and by Awwal, Griffin, and Scalise (2015) and Adams et al. (2015). Full linguistic analysis of communication was not performed, but rudimentary coding of instances of usage such as "why," "where," "how," and "?" was used to determine patterns of communication between partners. Actions were coded as dichotomous items based on presence or absence of a behavior. The frequency of occurrence of a behavior was used as a quasi-measure of difficulty (Adams et al., 2015) for calibration via item response modeling. Students' abilities were estimated with a proportional score of the maximum possible, based on the behaviors they had the opportunity to exhibit. If a student had the opportunity to display a specific behavior (such as responding to his or her partner's chat), he or she would be scored for that item. If the student did not have the opportunity (if, for example, his or her partner did not provide a resource for him or her), then that item was not scored and was considered as missing data. Complex algorithms with "if . . . then" propositions were utilized to create the indicators.

The following sequence of events (with indicators in parentheses) illustrates the scoring: Student A is prompted by text on the screen to type a number into a box, at which point Student B will see a different number on his or her screen. Student A is scored for entering a number (1) and for "chatting" to his or her partner before (2) and after (3) entering the number. Student B is scored for "chatting" to his or her partner before (5) and after (6) a number appears on his or her screen. If Student A does not enter a number, he or she is scored 0 for (1) but a *missing* for (3), and Student B receives a *missing* for (6). Chatting before the entering of a number—(2) and (5)—would be scored for each partner as that action is not dependent on Student A entering a number. In this way students are scored for their own contribution to the collaborative process.

Due to the design of the construction of indicators, there was a large amount of missing data in the data set. Missing data were considered as "missing at random" (Soley-Bori, 2013). This consideration was based on the notion that the missing data of a student was not a result of his or her own actions but was "random." In truth, most missing data were a result of the partner not proceeding with the activity and thus not entirely random. Nevertheless, the conception of "random" holds, as the student does not intentionally skip or miss indicators—the indicators are simply not presented. The use of Rasch analysis requires that dichotomous indicators be scored (0, 1) or not scored (missing). When a student does not have the opportunity to exhibit a behavior, the score must therefore be considered missing, as the other alternative—scoring the student as correct or incorrect on a behavior he or she could not have exhibited—would be folly. Missing data were not used to calculate a student's ability estimate; students were scored only on the indicators they had the opportunity to present.

During the calibration procedure, the indicators were pruned for consistency with the CPS construct. Some items were found not to correlate with CPS performance, despite the initial conceptualization of the action being an indicator of ability. For example, in the sequence described above, Student B chatting before a number appears (5) may not have corresponded with overall CPS ability. Indicators with low discrimination were not used in the calibration of student scores and were removed from the set of indicators using the procedure described by Harding and Griffin (2016).

The algorithms used for coding the indicators were either specific to each task or general to all tasks, and either unique to one student (A or B) or common to both students. If items were designed to be common to both students (which was necessary to have students placed on the same scale), Pearson's chi-square test was used to calculate the significance of any difference observed between aggregates of Student A's and Student B's scores. If there was a significant difference, the indicator was not considered as "common," because the difficulty parameter was not consistent. Those indicators were then scored separately for Student A and Student B.

In total, 88 indicators of CPS behavior were used over the four tasks. Survey questions were coded as polytomous depending on the number of categories (0, 1, 2, or 3). Categorization of the survey questions was based on the same theoretical framework used to categorize the indicators from the process data. Indicators that fit the model (in terms of the underlying construct) were used to calculate students' abilities.

### *Analysis*

The data were analyzed using the Rasch one-parameter model (Rasch, 1960/1980) adjusted for partial-credit coding. According to the model, the ability of each student and the difficulty of each indicator (difficulty measured by the frequency of occurrence of the behavior) governs the likelihood of the student scoring 1 on each dichotomous indicator. Masters extended the simple logistic model to the partial-credit model (Masters, 1982; Wright & Masters, 1982), allowing the analysis of polytomous items. Of the 88 indicators of CPS, 17 were coded as partial credit and 71 as dichotomous, resulting in a total of 114 parameters estimated. Student ability estimates, $\theta_n$, and item difficulty estimates,

$\delta_i$, were marginal maximum likelihood estimates obtained using an efficient maximum algorithm. The process of derivation of the student ability estimates was described by Adams et al. (2015). The ability parameter $\theta_n$ of person $n$, who had provided responses to a set of items or indicators $\Omega_n$, was estimated using an iterative process in which, after iteration $t$, student's ability is denoted $\theta_n^{(t)}$ as follows:

Let $k =$ possible score for item $i\left(0,1\ldots.m_i\right)$

Where $s =$ all possible categories $\left(0,1\ldots m_i\right)$

and $j =$ all possible categories $\left(0,1\ldots k_i\right)$

$$\text{Let } p_{nik}^{(t)} = \frac{e^{\left(k\theta_n^{(t)} - \sum_{j=0}^{k}\delta_{ij}\right)}}{\sum_{s=0}^{m_i} e^{\left(s\theta_n^{(t)} - \sum_{j=0}^{s}\delta_{ij}\right)}} \quad \left(\text{note; } e^{\sum_{k=0}^{0}(\theta_n - \delta_{ik})} \equiv 1 \text{ and } \delta_{i0} \equiv 0\right)$$

$$\text{Let } e_{ni}^{(t)} = \sum_{k=0}^{m_i} k p_{nik}^{(t)}$$

$$\text{Let } v_{ni}^{(t)} = \sum_{k=0}^{m_i} k^2 p_{nik}^{(t)} - \left(\sum_{k=0}^{m_i} k p_{nik}^{(t)}\right)^2$$

Let $r_n =$ score of student $n$ on items in set $\Omega_n$

Iterate over

$$\theta_n^{(t+1)} = \theta_n^{(t)} + \frac{r_n - \sum_{i\epsilon} e_{ni}^{(t)}}{\sum_{i\epsilon} v_{ni}^{(t)}} \text{ stop when } \left|\theta_n^{(t+1)} - \theta_n^{(t)}\right| < 0.001$$

Data were analyzed for a single latent dimension and for a two-dimensional construct consisting of social and cognitive dimensions. Between-item dimensionality was analyzed using the multidimensional random-coefficients multinomial logit model as described by Adams, Wilson, and Wang (1997). ConQuest computer software (Adams, Wu, & Wilson, 2012) was used to calibrate the item and person data.

For estimation of parameters, average indicator difficulty was arbitrarily set to zero, and student ability estimates were allowed to vary. The range of latent student ability estimates was compared to the range of indicator difficulties to check that the tasks were appropriately matched to students' abilities (Figure 7). For the two-dimensional model, both dimensions were plotted, with the total average indicator difficulty constrained to zero (logits). This allowed a visual representation of the differences in the difficulty estimates of the social and cognitive indicators, as shown in Figure 8. The correlation between dimensions was estimated and interpreted based on student variance shared between dimensions. The correlation estimate is effectively corrected for attenuation caused by measurement error (Adams et al., 2012).

### *Model Fit and Reliability*

Fit statistics were estimated as residual-based indices as described by Wu (1997), who extended those described by Wright and Masters (1982). Both unweighted and weighted fit were examined as evidence that the underpinning construct was represented by the indicators. Weighted fit is the mean-squared difference between the observed and the estimated difficulty of each score, weighted by the variance of the assigned score, referred to as INFIT (information-weighted mean-squared residual goodness-of-fit statistic). Unweighted fit is outlier sensitive and based on traditional chi-square statistics, referred to as OUTFIT (outlier-sensitive mean-squared residual goodness-of-fit statistic). According to Linacre (2002), INFIT reports overfit for Guttman patterns and underfit for alternative curricula or idiosyncratic groups, and OUTFIT is more sensitive to responses to indicators with difficulty far from a person's ability and vice versa.

If the model fits the data, then the INFIT and the OUTFIT should approximate to 1. Both fit statistics are sensitive to large samples, and the confidence interval will narrow as the sample size increases. Acceptable fit is often quoted as ranging between 0.77 and 1.20 (as in Adams & Khoo, 1995). However, with the large sample of students involved in this study, a more acceptable range of fit was considered to be between 0.8 and 1.2. Indicators conforming to these criteria were retained for analysis.

Reliability estimates for indicator and student separation were identified using ConQuest (Adams et al., 2012).

### *Validity*

To examine threats to consequential validity, evidence of fairness between the two different test versions (Student A and Student B roles) and lack of bias between versions of the assessments for each participating country will be presented below. Evidence of the potential benefits for teaching and learning will be considered in the Discussion.

Indicators (items) common to both students (A and B) were used to link the student assessments. Common indicators were tested for differences in parameter difficulty (tested for differential item functioning [DIF]), and only those retaining identical difficulties for Student A and Student B were considered as "common" indicators. Common person equating methods were used to establish item difficulty parameters that were comparable across tasks.

Of the 3,004 students completing the tasks, 2,944 maintained their role as Student A or Student B for all tasks taken. That is, just 60 students changed from Role A to Role B or from Role B to Role A for different tasks during the

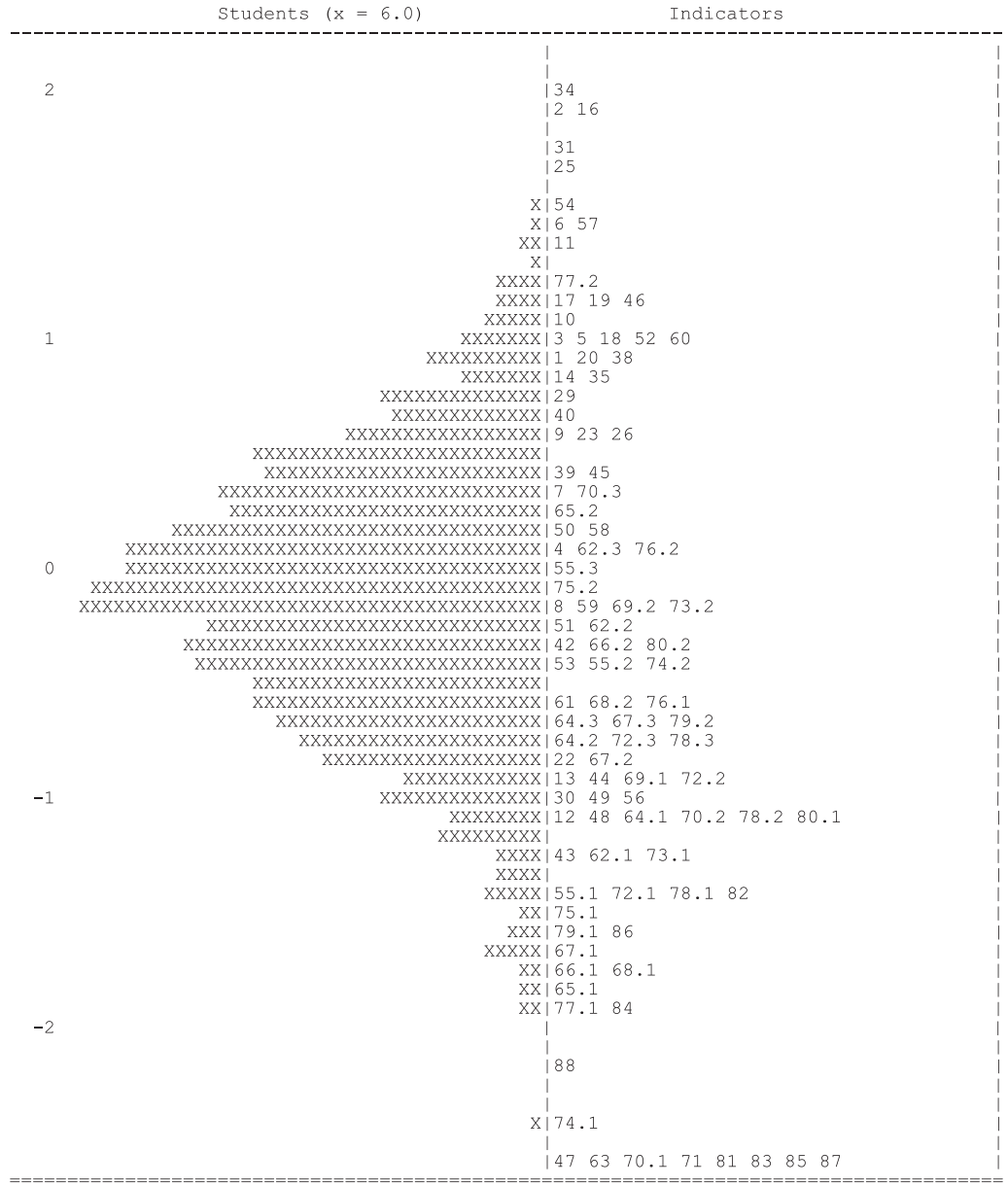


FIGURE 7. *Variable map of one-dimensional analysis.*

assessment—perhaps a result of the students not being told that they were able to swap roles. Student A was scored for an overall item pool that differs from that of Student B, other than for common items. This was accounted for by the model, and there was no hypothesized advantage or disadvantage based on the role the student played in the assessment. The difference in mean latent ability was determined for the group of students who maintained the Student A role and the group who maintained the Student B role. This was achieved by running a general facet analysis with role (A or B) as a main effect.

To determine that indicator functioning was unaffected by language-, culture-, or country-specific factors, the within-country indicator difficulty estimates were compared between countries. This was performed as a DIF comparison by correlating the difficulties of the parameter estimates in each country.

The main effects of country differences were also compared. This was achieved by running a general facet analysis with country as a main effect. Importantly, unlike the main effects of differences in ability estimates per role, the hypothesis was that there were likely to be differences in the averages of the ability estimates per country, because the students from each country were selected by opportunistic sampling and there could be real differences between countries in CPS abilities.

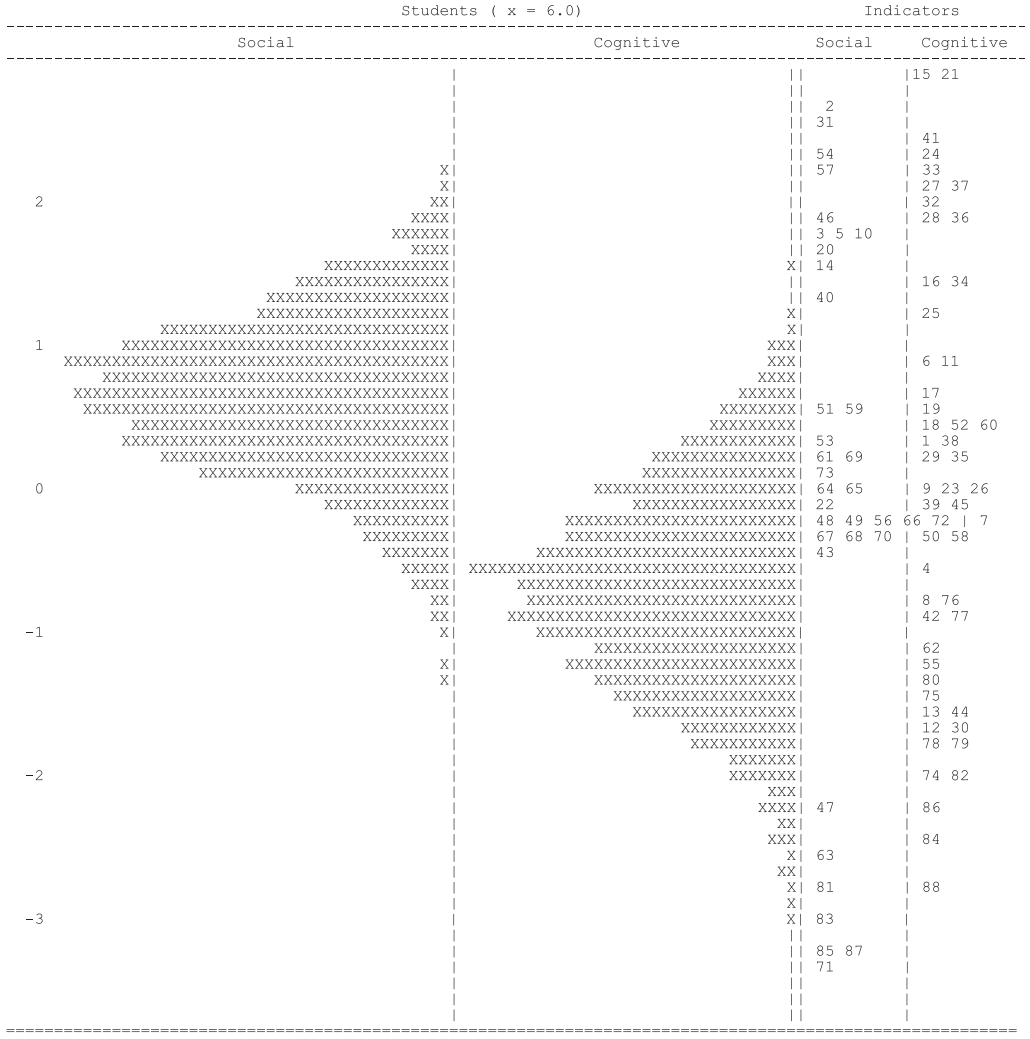


FIGURE 8. *Variable map of two-dimensional analysis.*

### *Dependencies*

An analysis of dependencies between students skills on the social and cognitive dimensions was conducted to detect potential interactions that may have resulted from the framing of the CPS tasks in a mathematical context, that is, to detect the potential influence of mathematical ability on the results. Correlations were calculated in SPSS with the square of the correlation coefficient ($R^2$) reported as the percentage of the variation in one variable that is related to the variation in the other. An $F$ test for significance was conducted.

To investigate the dependency of one partner's social or cognitive CPS skills on the other, Student A ability estimates were correlated with Student B ability estimates on both the social and cognitive dimensions.

Other possible dependences within the tasks, dependencies across tasks, temporal dependencies, and dependencies across people have not yet been investigated.

## Results

### *Analysis*

*One-dimensional model.* Calibration of the four tasks included 88 indicators: 27 unique for Student A, 20 unique for Student B, and 41 common for Students A and B. Indicator parameter estimates are summarized in Table 2; minimum and maximum item parameter estimates are shown for each of the four tasks, with mean estimates reported. The mean of the total item difficulty estimates was 0.000, as specified by the model. Fit statistics are summarized with mean INFIT, mean OUTFIT, and variance of these reported (Table 2). The total mean-square weighted estimate for the joint calibration averaged 1.003 with a variance of 0.003, indicating that the model fit the data. The items were well separated with an item separation reliability of 0.999; expected a posteriori/plausible value (EAP/PV) person separation was adequate at 0.739. Warm's weighted likelihood

TABLE 2
*Summary Statistics: One-Dimensional Model*

| | Item Parameter Estimate | | | Fit Statistics | | | |
|---|---|---|---|---|---|---|---|
| Task | Mean | Min | Max | Mean INFIT | Variance of INFIT | Mean OUTFIT | Variance of OUTFIT |
| Hexagons | 0.154 | −3.601 | 1.947 | 0.994 | .002 | 1.012 | .008 |
| Warehouse | 0.604 | −3.847 | 3.661 | 1.018 | .002 | 1.004 | .016 |
| Game of 20 | 1.060 | −4.081 | 2.872 | 1.019 | .002 | 1.063 | .009 |
| Small Pyramids | −0.611 | −4.046 | 1.552 | 0.995 | .003 | 1.013 | .013 |
| Survey | −1.105 | −4.165 | −0.270 | 0.992 | .008 | 0.984 | .021 |

*Note.* INFIT = information-weighted mean-squared residual goodness-of-fit statistic; OUTFIT = outlier-sensitive mean-squared residual goodness-of-fit statistic.

TABLE 3
*Summary Statistics: Two-Dimensional Model*

| | Item Parameter Estimate | | | Fit Statistics | | | |
|---|---|---|---|---|---|---|---|
| Task | Mean | Min | Max | Mean INFIT | Variance of INFIT | Mean OUTFIT | Variance of OUTFIT |
| Hexagons | .088 | −2.787 | 2.715 | 1.002 | .002 | 1.002 | .007 |
| Warehouse | .447 | −2.542 | 3.173 | 1.026 | .003 | 1.019 | .024 |
| Game of 20 | .767 | −3.259 | 2.446 | 1.031 | .003 | 1.100 | .013 |
| Small Pyramids | −.380 | −3.222 | 2.335 | 1.000 | .005 | 1.015 | .014 |
| Survey | −.881 | −3.348 | 0.223 | 0.985 | .005 | 0.976 | .012 |

*Note.* INFIT = information-weighted mean-squared residual goodness-of-fit statistic; OUTFIT = outlier-sensitive mean-squared residual goodness-of-fit statistic.

estimate reliability was 0.617 and maximum likelihood estimate reliability was 0.644, indicating a satisfactory estimate of person's ability based on the assessments.

The mean of the latent ability distribution for all four tasks was −0.163 logits (*SE* = 0.011), indicating that the task indicators were well matched to students' abilities. This is demonstrated visually in the variable map shown in Figure 7.

*Two-dimensional model.* The data set was then calibrated based on the assumption of two logical dimensions, social and cognitive, as described in Table 1 (Hesse et al., 2015). Indicators were designated as belonging to the social (36) or cognitive (52) dimension. Indicator parameter estimates are summarized in Table 3. Minimum and maximum item parameter estimates are shown for each of the four tasks, based on analysis using the two dimensions, with mean estimates reported. Fit statistics are summarized with mean INFIT, mean OUTFIT, and variance of these reported (Table 3). The total mean-square weighted estimate for the joint calibration averaged 1.009 with a variance of 0.004, indicating that the model fit the data. There were no substantial fit differences between the one- and two-dimensional models, providing evidence that CPS can be treated as a unitary construct or as a construct with social and cognitive dimensions. The items were well separated with a separation reliability of 0.997; EAP/PV person separation was 0.685 on the social dimension and 0.714 on the cognitive dimension. This separation indicates that the items on the cognitive dimension were marginally more accurate for determining students' ability.

The means of the total item parameter estimates for each of the two dimensions were set to 0.000, allowing the students' abilities, as measured for each dimension, to vary. On the social dimension, students' mean latent ability distribution was 0.634 logits (*SE* = 0.011), and on the cognitive dimension, students' mean latent ability distribution was −0.734 logits (*SE* = 0.014). The students scored higher on the indicators assigned to the social dimension than on the indicators assigned to the cognitive dimension. This is represented visually in the variable map shown in Figure 8.

The dimensions were highly correlated ($r$ = .83), with 69% of student variance shared between the dimensions. Thus, students who scored highly on one dimension were likely to score highly on the other.

TABLE 4
*Student Role as Facet*

| Student Role | Estimate | Error | OUTFIT Mean | OUTFIT Confidence Interval | OUTFIT *T* | INFIT Mean | INFIT Confidence Interval | INFIT *T* |
|---|---|---|---|---|---|---|---|---|
| A | .001 | .006 | 1.05 | [0.93, 1.07] | 1.3 | 1.02 | [0.92, 1.08] | 0.5 |
| B | −.001 | .006 | 1.03 | [0.93, 1.07] | 0.8 | 1.02 | [0.92, 1.08] | 0.4 |

*Note*. INFIT = information-weighted mean-squared residual goodness-of-fit statistic; OUTFIT = outlier-sensitive mean-squared residual goodness-of-fit statistic.

### *Validity*

*Student role as a facet.* Data collected from the 2,944 students who maintained either Role A or Role B for the combination of the four tasks they completed were analyzed to determine differences in mean latent ability estimates by role. The estimates were hypothesized to be equal for the two roles due to the sampling methods involved. The testing population of students from each country was selected on the basis of teachers and schools wanting to use the assessments—a nonprobabilistic sampling method. The population of students who were appointed Role A or Role B was, in contrast, a type of stratified random sampling, where students from a single class, in a single school, from each of the countries involved in the study, were separated into dyads. This method created a homogeneous subgroup from which an equal sample of Student A and Student B roles were drawn. Therefore, on an aggregate basis, the sample of Student A students was roughly matched with the sample of Student B students for year/grade level, school, district, culture, language, and teacher-specific differences, even though teachers were not asked to separate students based on these characteristics.

Role was included as a facet in the measurement construct, and main effects were examined. The parameter estimates for Role A and Role B differ by just 0.002 logits, which is less than the standard error of this estimate (0.006); therefore, there was no effective difference in the mean latent ability estimate based on role, as shown in Table 4.

*Country as a facet.* To examine the possibility of DIF between the student samples in the different countries involved in this study, data were calibrated separately for each country and item parameter estimates were compared. The correlation between item parameter estimates for each country was used as an indication of the amount of DIF between countries. All the countries were compared and correlations (r) are shown in Table 5. All correlations were significant (two-tailed significance 0.000), demonstrating no major differences in the way the indicators were measuring CPS in the different countries. This suggests that students follow the same process when solving mathematical tasks collaboratively, regardless of culture, language, or country of origin. The criteria for selecting indicators have provided a set of invariant item difficulties across countries. Thus, participating groups of students can be compared using the set of indicators presented.

TABLE 5
*Country Item Parameter Correlations (r)*

| Country | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | — | .902 | .894 | .918 | .965 | .896 |
| 2 | .902 | — | .924 | .896 | .919 | .918 |
| 3 | .894 | .924 | — | .894 | .907 | .917 |
| 4 | .918 | .896 | .894 | — | .936 | .934 |
| 5 | .965 | .919 | .907 | .936 | — | .922 |
| 6 | .896 | .918 | .917 | .934 | .922 | — |

Country was added as a facet to the calibration of the joint data set (2,944 cases) to determine whether students from each country differed in terms of their ability estimates. This comparison can be made because the indicator difficulties per country are invariant, as shown in Table 5. The data collection was focused on the psychometric properties of the tasks, and therefore countries have been deidentified. Countries differed in their implementation of the tasks. Therefore differences in ability cannot be used to generalize to the population. Rather, this analysis demonstrates that differences in abilities between participating groups, whether matched on age, nationality, gender, or other student background characteristics, can be estimated using these tasks, and generalization to national means would be appropriate if a proper probability sample had been drawn. In contrast to using student role as a facet, it was hypothesized that differences in main effects would be identified due to country differences. The difference between the highest-performing country group (Country 3) and the lowest-performing country group (Country 4) is 0.71 logits, which is above the parameter estimate errors, as shown in Table 6. This demonstrates that student groups can be compared in terms of ability using the set of tasks presented in this study. The observation that the main effects of country and student role (A or B) have no interaction with the indicators provides confidence that such studies can be undertaken without fear of bias.

TABLE 6
*Country as Facet*

| | | | OUTFIT | | | INFIT | | |
|---|---|---|---|---|---|---|---|---|
| Country | Estimate | Error | Mean | Confidence Interval | *T* | Mean | Confidence Interval | *T* |
| 1 | .166 | .012 | 1.12 | [0.88, 1.12] | 1.9 | 1.08 | [0.86, 1.14] | 1.1 |
| 2 | −.062 | .013 | 0.93 | [0.85, 1.15] | −0.9 | 0.95 | [0.84, 1.16] | −0.6 |
| 3 | .407 | .013 | 1.02 | [0.84, 1.16] | 0.3 | 1.01 | [0.83, 1.17] | 0.2 |
| 4 | −.303 | .009 | 1.06 | [0.91, 1.09] | 1.3 | 1.04 | [0.9, 1.1] | 0.7 |
| 5 | −.111 | .011 | 1.01 | [0.88, 1.12] | 0.2 | 0.96 | [0.87, 1.13] | −0.6 |
| 6 | −.097 | .026 | 0.97 | [0.84, 1.16] | −0.3 | 0.97 | [0.82, 1.18] | −0.3 |

*Note.* INFIT = information-weighted mean-squared residual goodness-of-fit statistic; OUTFIT = outlier-sensitive mean-squared residual goodness-of-fit statistic.

### *Dependencies*

To address Research Questions 3 and 4, Pearson product-moment correlation coefficients were computed to assess the relationship between variables. In each case, the relationship was positive. However, the strength of the relationship varied, as was expected. Student A's social ability estimates were weakly associated with partner Student B's social ability estimates, $r = .341$, with 11.6% of the variance in scores explained by the performance of the partner ($p = .000$). Student A's cognitive ability estimates were weakly associated with partner Student B's cognitive ability estimates, $r = .336$, with 11.3% of the variance in scores explained by the performance of the partner ($p = .000$) (Figure 9, A and B).

In a comparison of CPS performance on mathematics tasks and content-free tasks, students' social ability estimates were strongly correlated, $r = .819$, with 67% variance shared between social estimates using different task types ($p = .000$). Students' cognitive ability estimates were moderately correlated, $r = .692$, with 47.9% shared between cognitive estimates using different task types ($p = .000$) (Figure 9, C and D).

## Discussion

The Rasch model fit to the data provided evidence that the assessment is measuring a consistent latent trait, with mean item INFIT ranging from 0.992 to 1.019 (Table 2) for the four mathematics tasks and the survey using the one-parameter model. The two-dimensional model (considering social and cognitive dimensions) also fit the data, with mean item INFIT ranging from 0.985 to 1.031 (Table 3). The person and item separation reliabilities were adequate, and the assessments provided a satisfactory estimate of person's ability. The data presented in this article show clear evidence that individual CPS ability can be measured using content-dependent mathematical tasks, and the tasks were well matched to student abilities.

Various aspects of assessment validity have been addressed. Consequential validation as defined by Messick (1995) combines evidence that the performance assessments have potential benefits for teaching and learning with evidence that adverse consequences are minimal. Messick outlines the need to evaluate the intended and unintended consequences of score interpretation, particularly in association with bias in scoring and interpretation and with unfairness in test use. In terms of bias, students from six countries were tested, and even though the abilities of these opportunistic samples were different from one another (Table 6), the mean correlation of the parameter estimates (that is, the order of difficulty of the indicators in the tasks for each country) was 0.92 (Table 5), indicating that the construct measured did not differ from country to country. This finding was also reported when the entire host of CPS tasks (both content dependent and content free) was calibrated and reported by Harding and Griffin (2016). The mathematics tasks can therefore be used in the six countries tested without compromising the validity of the assessment, although much work remains to be done to ensure that the mathematics CPS assessments meet other criteria for a "fair" test, so that students are not advantaged or disadvantaged by linguistic, communicative, cognitive, cultural, physical, or other characteristics (AERA et al., 2014).

The results of the facet analysis with student role as a factor (Table 4) showed that there was no difference in the mean latent ability estimate based on role, which was inevitable given the design and use of Rasch analysis to estimate the difficulty of indicators and to score students accordingly. Given the asymmetric nature of the tasks and indicators, it was vital to provide evidence that there was no advantage or disadvantage to the role that the student took in the task. The data showed that a student's score was not dependent on his or her role, or on the set of indicators he or she was scored on, and that the assessments are "fair" for each partner.

The influence of the ability of one student on the ability estimate of the other was investigated within the constraints

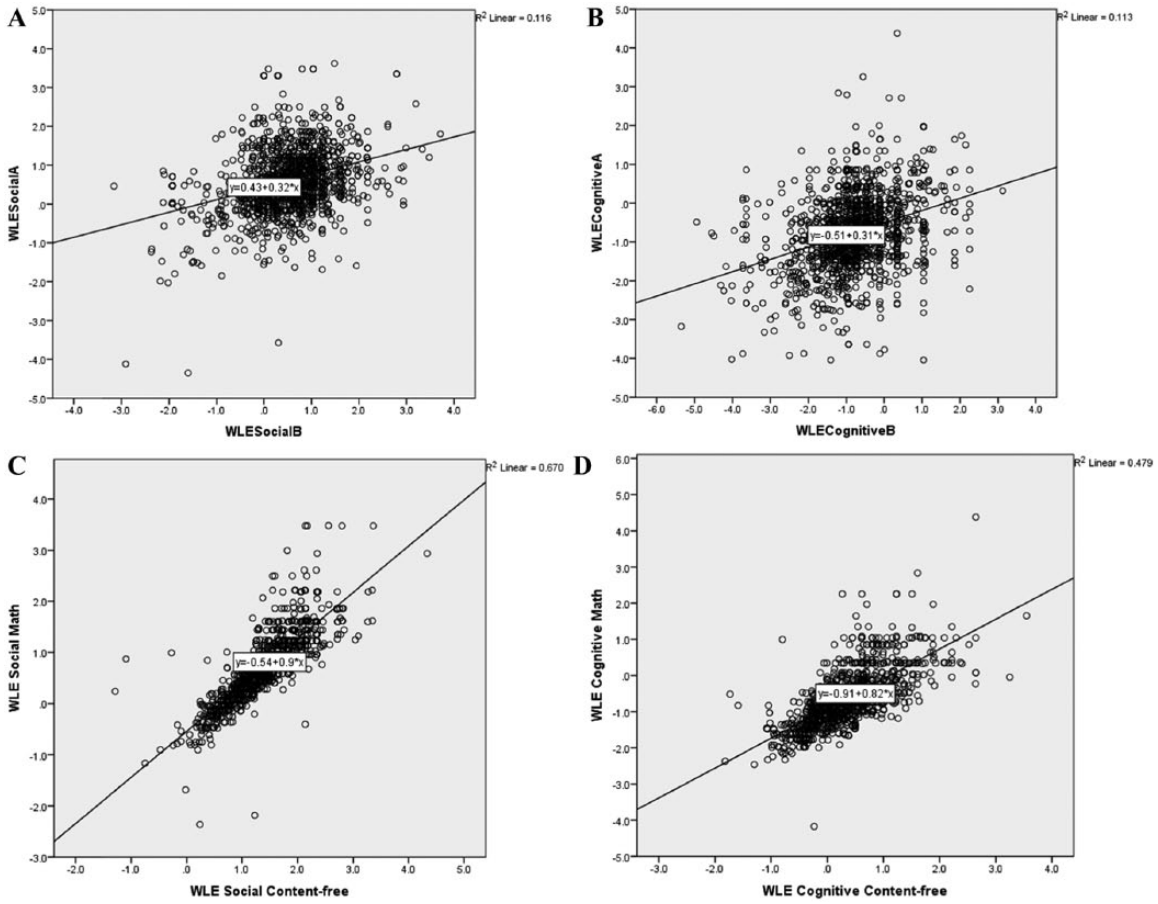


FIGURE 9. *Correlation studies of dependencies. (A) Correlation of social ability estimates on math tasks; Student A versus Student B. (B) Correlation of cognitive ability estimates on math tasks; Student A versus Student B. (C) Correlation of social ability estimates on math tasks versus content-free tasks. (D) Correlation of cognitive ability estimates on math tasks versus content free-tasks.*

of the sampling technique to address Research Question 3. It was an intention of the assessment that the ability estimates be independent for each partner. The indicators were carefully designed to avoid dependence, with each student having the opportunity to exhibit skills of CPS without reliance on what the partner says or does. There are likely dependencies existing regardless, perhaps in a psychological manner, where one enthusiastic or particularly knowledgeable partner is able to guide the other partner through the assessment. This is a conflicting situation where, ideally, students will be guiding each other through the tasks via collaboration, but the score of each individual student should not be influenced by the partner's CPS ability. Partners' ability estimates should not be correlated to maintain discriminant validity of the assessment. In other words, students may influence one another, but their scores should not be dependent, as each student was judged on his or her unique set of CPS skills.

The mechanisms for dealing with potential dependencies in the data was not the focus of the ATC21S project, from which the data used in this paper were drawn. Hence the sampling distribution of students was not random, and any inferences from modeled dependencies need to be taken with caution. Artificial inflation of dependencies between partners will be evident, as student pairings took place within class at school level. A "matching" of student abilities consequently took place, as students within the same class in the same year/grade level and in the same country will be more homogenous compared to other cohorts. This is a major limitation in the research design. Further examination of the potential dependency of one partner on another should be undertaken with data drawn from an appropriate randomized study. Despite the suboptimal sampling conditions, only 11.6% and 11.3% of student variation in social and cognitive skills, respectively, were shared between partners. This finding suggests that the sole source of dependencies could be matched pairing as a by-product of sampling in pairs by class. Unfortunately, there is no way to separate the effect of the pairing on the effect of any real dependencies within the assessment without either (a) an external basis, such as separate mathematics and collaboration assessments for those same students, to determine the base level of ability relationship between students or (b) a proper randomized study

allocating partners at random across schools, year/grade levels, and countries.

The Rasch simple logistic model was used to assess unidimensionality, response category functioning, item fit, personal reliability, and item invariance across national samples, language, and role (A or B) of the student. An important specification of the Rasch model is that local independence is maintained. If responses to items (or in this case, behaviors manifested) are dependent upon other persons or items, then the Rasch model is not appropriate for ability estimation. Considering the likelihood of possible dependencies remaining (factors other than those accounted for by ability on the latent trait), different types of analysis should be considered in future research either to discount remaining dependencies in the data or to model them. Concerns regarding loss of local independence have been discussed recently by Griffin (2017, p. 128).

The mathematical tasks used to measure CPS skills were designed to be useful to teachers for reporting students' level of development of CPS skills on both the social and cognitive dimensions while students practice or learn mathematical skills in the classroom. Even though "correct answer" was not the focus of the assessments, elements such as systematicity, relationships, contingencies/rules, and hypotheses were likely to correlate highly with general mathematical ability. The focus of the assessment was on the CPS process rather than mathematics; however, the mathematical ability of the student was hypothesized to influence his or her score on the cognitive or "problem-solving" part of the CPS construct. Students' social skills were hypothesized to be more highly correlated between task types than cognitive CPS skills; the results confirmed this hypothesis, providing evidence for convergent validity of the assessments. The main limitation of the comparison between mathematics and content-free tasks was that students in the sample did not complete a large number of nonmathematics tasks, producing noise in the data and reducing the real correlation of abilities on both the social and cognitive dimensions.

When social skills were compared, there was 67% shared variance in student ability estimates between mathematics and content-free tasks, suggesting that the type of task did not strongly affect the measurement of the students' social CPS skills. However, when cognitive skills were compared, there was only 47.9% shared variance in student ability estimates between mathematics and content-free tasks. This indicates that the cognitive ability estimates were impacted by the content of the tasks (students scored differently when the tasks were framed in the mathematics context), but this does not imply that the measure is inaccurate. There is a hypothesized relationship between mathematical understanding and skill and the "problem-solving" or cognitive part of the CPS construct. As more is understood about the construct of CPS and how it relates to other content areas, such as mathematics, the scope to improve CPS assessments widens. How that improvement will manifest is unknown. If the cognitive component of CPS is in fact correlated with mathematical ability, then it is not necessarily the case that assessments should be designed to avoid this relationship. Teachers using CPS assessments that are influenced by particular content abilities should be made aware of the influences and educated on how to interpret the results.

Many research groups and governing bodies have avoided the issue of dependencies between partners by creating human-to-computer agent tasks (Rosen, 2017; von Davier, 2017). Other approaches, such as modeling both group and individual CPS skill dynamics and partitioning out the effect of the group from the ability of the individual, are suggested (von Davier & Halpin, 2013). This study examined an assessment approach that attempted to avoid dependence between partners by assessing only behaviors each individual student had the opportunity to present. There are many ways to approach the assessment of CPS, and the findings of this study do not seek to resolve considerations by arguing for or against a particular type of assessment or analysis of the assessment. Rather, the aim was to contribute to shared knowledge on the basis of creating, using, and analyzing assessments for CPS, particularly focusing on those that would be useful for teachers.

This paper contributes to a growing body of evidence that problem solving, whether individual or collaborative, is not only a mathematics-based skill. With other studies of its kind, it raises the possibility that tasks designed to measure student collaboration can be constructed with any curriculum base or in a context that is not linked to a particular curriculum component. CPS tasks can be linked to the subject areas of mathematics and thus alleviate pressure on teachers to measure this transverse skill without diverging from the subject in the classroom.

## Acknowledgments

This research was funded in part by ARC-SRI: Science of Learning Research Centre (Project No. SR1203000015) Australia, as a secondary data analysis from the ATC21S (Assessment and Teaching of 21st Century Skills) project conducted 2009 to 2012.

## References

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

Adams, R. J., & Khoo, S. T. (1995). *Quest: An interactive item analysis program*. Melbourne, Australia: Australian Council for Educational Research.

Adams, R. J., Wilson, M., & Wang, W.C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.

Adams, R. J., Wu, M. L., & Wilson, M. R. (2012). ACER ConQuest 3.0 [Computer program]. Melbourne: ACER.

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

Awwal, N., Griffin, P., & Scalise, S. (2015). Platforms for delivery of collaborative tasks. In P. Griffin, & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 105–114) Dordrecht, Netherlands: Springer.

Care, E., Anderson, K., & Kim, H. (2016). *Visualizing the breadth of skills movement across education systems*. Retrieved from http://skills.brookings.edu/

Care, E., Griffin, P., Scoular, C., Awwal, N., & Zoanetti, N. (2015). Collaborative problem solving tasks. In P. Griffin, & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 85–94) Dordrecht, Netherlands: Springer.

Economist Intelligence Unit. (2015). *Driving the skills agenda: Preparing students for the future*. Retrieved from https://www.eiuperspectives.economist.com/sites/default/files/Drivingtheskillsagenda.pdf

Greiff, S., Holt, D., & Funke, J. (2013). Perspectives on problem solving in educational assessment: Analytical, interactive, and collaborative problem solving. *Journal of Problem Solving*, *5*, 71–91.

Griffin, P. (2014, October). *The changing nature of education, schools and employment readiness*. Paper presented at the Korea Development Institute Workshop on Human Capital Policy, Seoul, South Korea. Retrieved from: http://www.kdi.re.kr/upload/10152/Paper_12.pdf

Griffin, P. (2017). Assessing and teaching 21st century skills: Collaborative problem solving as a case study. In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 113–134). Cham, Switzerland: Springer International.

Griffin, P., & Care, E. (2015). *Assessment and teaching of 21st century skills: Methods and approach*. Dordrecht, Netherlands: Springer.

Griffin, P., Care., E., & Harding, S. (2015). Task characteristics and calibration. In P. Griffin, & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 133–177). Dordrecht, Netherlands: Springer.

Griffin, P., Care, E., & McGaw, B. (2012). The changing role of education and schools. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 1–15). Dordrecht, Netherlands: Springer.

Harding, S., & Griffin, P. (2016). Rasch measurement of collaborative problem solving in an online environment. *Journal of Applied Measurement*, *17*(1), 35–53.

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). Dordrecht, Netherlands: Springer.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues & Practice*, *14*(4), 5.

Ministry of Education, Singapore. (2015). *21st century competencies*. Retrieved from https://www.moe.gov.sg/education/education-system/21st-century-competencies

O'Neil, H. F. (1999). Perspectives on computer-based performance assessment of problem solving, *Computers in Human Behaviour*, *15*(3/4), 225–268.

O'Neil, H. F., Chuang, S., & Chung, G. K. W. K. (2003). Issues in the computer-based assessment of collaborative problem solving. *Assessment in Education: Principles, Policy and Practice*, *10*, 361–373.

Organisation for Economic Co-operation and Development. (2015). *OECD skills outlook 2015: Youth, skills and employability*. Paris, France: Author. http://dx.doi.org/10.1787/9789264234178-en

Polya, G. (1973). *How to solve it*. Princeton, NJ: Princeton University Press. (Original work published in 1945)

Programme for International Student Assessment. (2010). *Field trial problem solving framework*. Retrieved from http://www.oecd.org/pisa/pisaproducts/46962005.pdf

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen,. Chicago, IL: University of Chicago Press. (Original work published 1960)

Rosen, Y. (2017). Assessing Students in human-to-agent settings to inform collaborative problem-solving learning. *Journal of Educational Measurement*, *1*, 36. doi:10.1111/jedm.12131

Soley-Bori, M. (2013). *Dealing with missing data: Key assumptions and methods for applied analysis*. Technical report, Boston University, Boston, MA.

von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Education Measurement*, *54*(1), 3–11.

von Davier, A. A., & Halpin, P. F. (2013). *Collaborative problem solving assessments and the assessment of cognitive skills: Psychometrical considerations* (Research Report 12-41). Princeton, NJ: Educational Testing Service.

von Davier, A. A., Zhu, M., & Kyllonen, P. C. (Eds.). (2017). *Innovative assessment of collaboration*. Cham, Switzerland: Springer International.

World Economic Forum. (2016). *What are the 21st-century skills every student needs?* Retrieved from https://www.weforum.org/agenda/2016/03/21st-century-skills-future-jobs-students/

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: Mesa Press.

Wu, M. L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalised item response models*. Unpublished Master's Dissertation, University of Melbourne, Australia.

## Authors

SUSAN-MARIE E. HARDING is a research fellow at the Assessment Research Centre, University of Melbourne. She specializes in quantitative research, including the design, development, and measurement of data collection tools and assessments. Her focus is generating and addressing key assessment questions using Rasch psychometric analysis and other methods.

PATRICK E. GRIFFIN is chief investigator at the Assessment Research Centre, University of Melbourne. He is the founding director and emeritus professor of the Assessment Research Centre.

He has published widely on assessment topics that include the development and calibration of instruments to measure collaborative problem solving and other 21st century skills as well as literacy, numeracy, and problem-solving proficiency.

NAFISA AWWAL is a research fellow at the Assessment Research Centre, University of Melbourne. She specializes in the design, development, and implementation of Web-based educational assessments, reporting, and other data collection tools. Her research includes data management, analysis, and item writing.

BM M. ALOM is a programmer at the Assessment Research Centre, University of Melbourne. He works on software development specifically for the assessment of 21st century skills, including collaborative problem solving. He programs online assessment tasks and scoring algorithms.

CLAIRE SCOULAR is a research fellow at the Assessment Research Centre, University of Melbourne. Her work includes the design, implementation, interpretation, and analysis of educational assessment and psychological measurement.