

*International Journal of Learning, Teaching and Educational Research*  
Vol. 17, No. 2, pp. 138-158, February 2018  
<https://doi.org/10.26803/ijlter.17.2.9>

## Assessing Online Collaborative Problem Solving Among School Children in Finland: A Case Study Using ATC21S™ in a National Context

**Arto K. Ahonen**

Finnish Institute for Educational Research  
University of Jyväskylä  
Jyväskylä, Finland

**Susan-Marie Harding**

Assessment Research Centre  
University of Melbourne  
Melbourne, Australia

**Abstract.** Online collaborative problem-solving tasks were piloted in five Finnish schools participating in the Assessment and Teaching of Twenty-first Century Skills (ATC21S™) project. Process stream data from online tasks were captured from 480 Finnish students who explored dyad problem spaces. The log file data were explored to identify indicative behaviours of collaborative problem solving (CPS), which were coded into a series of dichotomous indicators. The Rasch simple logistic model was used to analyse student performance across the social and cognitive dimensions of CPS. Analysis of the Finnish data indicated that the construct can be interpreted using a two-dimensional model, matching the findings of the total cohort of 4,078 students. Further analysis of the data provided evidence that gender has no statistical impact on Finnish students' CPS abilities in the social or cognitive dimensions. Analysis of the differences in student abilities across grades within and between schools suggested that upper-year students had stronger CPS skills.

**Keywords:** Assessment, collaboration, problem solving, school, students.

### Introduction

#### *Twenty-first Century Skills*

Success in life and work in today's knowledge society calls for transversal and generic twenty-first century skills or key competencies, such as skills for learning, creative and critical thinking, collaboration, communication, civic skills and the ability to utilise information and communication technologies (ICTs)

(Ananiadou & Carlo, 2009; Binkley et al., 2012; Gordon et al., 2009; Pepper, 2011). Embedding twenty-first century skills in schools requires assessment systems that can measure rich learning accurately and complex tasks that can be used flexibly without overburdening students or teachers in the learning process (Rotherham & Willingham, 2010).

While the number of studies on twenty-first century skills is increasing, rigorous research in this field is still in its infancy (Greiff, Niepel, & Wistenberg, 2015). Few novel, technology-enhanced formative assessment practices, tools or systems reflecting generic twenty-first century skills have been studied. Also, current assessment practices do not always consider the richness and complexity of technology-enhanced collaboration. Even though collaboration is regarded as one of the most crucial skills for future learning (Silva, 2009) and is already a common part of today's learning environments, the assessment practices for collaborative learning remain relatively vague (Strijbos, 2011).

The Assessment and Teaching of Twenty-first Century Skills (ATC21S™) project explores new ways to assess twenty-first century skills and link them to teaching interventions designed to deepen learning and move students to higher skill levels (see Figure 3) (Csapo, Ainley, Bennett, Latour, & Law, 2012; Griffin, Care, & McGaw, 2012). A computer-based assessment portal that includes tasks to assess collaborative problem solving (CPS) has been developed at the Assessment Research Centre at the University of Melbourne. International focus on the enhancement of twenty-first century skills in learning has also brought about the need for and interest in developing tools and methods for teaching and assessing these skills. To this aim, Finland joined in the ATC21S™ project as a founder country in 2009, the other participating countries were Australia, Costa-Rica, the Netherlands, Singapore and the USA (Ahonen & Kankaanranta, 2015).

## **Definition and Assessment of CPS**

CPS, a specific type of collaboration, has received increasing attention as an important twenty-first century skill suitable for assessment. Following the PISA 2015 study, CPS competency was described as the capacity of an individual to engage effectively in a shared activity, in which participants combine knowledge, efforts and skills to reach a common goal (OECD, 2017). In short, collaborative problem solving is a joint activity between dyads or small groups who seek to transform a current problem state into a desired goal state (Hesse, Care, Buder, Sassenberg, & Griffin, 2015). CPS is organised via directly observable, verbal and nonverbal signals. To succeed, participants need to communicate, exchange, and share in the process of identifying the parts of the problem, interpreting the connections between the parts and the relationships between action and effect (i.e., the rules) and proposing generalisations for a shared solution (Hesse et al., 2015). This 'side effect' of externalisation makes CPS a visible and measurable activity not only for individual problem solving

but also as a teachable skill (Hesse et al., 2015). Simply put, this measurable CPS competency is the capacity of an individual learner to engage effectively in shared problem solving.

CPS is not a uniform process, but rather a complex, coordinated activity among two or more individuals. CPS, particularly for ill-defined problems, cannot be scripted; learners must account for several factors that depend on situational affordances (Hesse et al., 2015). Hesse et al. (2015) identified five broad strands (see Figure 1) to define the CPS construct: (1) participation (readiness to share information and externalise thoughts), (2) perspective taking (the ability to take the others' perspectives into account), (3) social regulation (awareness of the strengths and weaknesses of group members), (4) task regulation (planning and monitoring skills to develop strategies for problem solving and shared problem representation), and (5) knowledge building (the ability to learn and build knowledge through group interaction). Social skills, which refer to the 'collaborative' part of problem solving, are about managing participants (including oneself), whereas cognitive skills (the 'problem solving' part) are about managing the task at hand (Hesse et al., 2015).

Shared regulation is a substantial requirement of successful collaboration in computer-supported collaborative learning (Järvelä & Hadwin, 2013; Miller & Hadwin, 2015). To investigate this concept from the angle of learning assessments, this study evaluates the described five-strand skills of students by measuring their performance of ATC21S™ tasks. Specific indicators relating to CPS ability have been designed based on and mapped to each of the skills and sub-skills described by Hesse et al. (2015). One hundred-fifty indicators used to analyse student CPS skills have been distributed across the five strands, and different modes of measurement have been applied to each strand (Harding & Griffin, 2016). Multiple research articles have been published describing the creation, coding and calibration of the ATC21S™ tasks, including those tasks specific to particular subject areas i.e. mathematics (Harding, Griffin, Awwal, Alom, & Scoular, 2017), however this article is the first to focus on gender, grade and school differences observed.

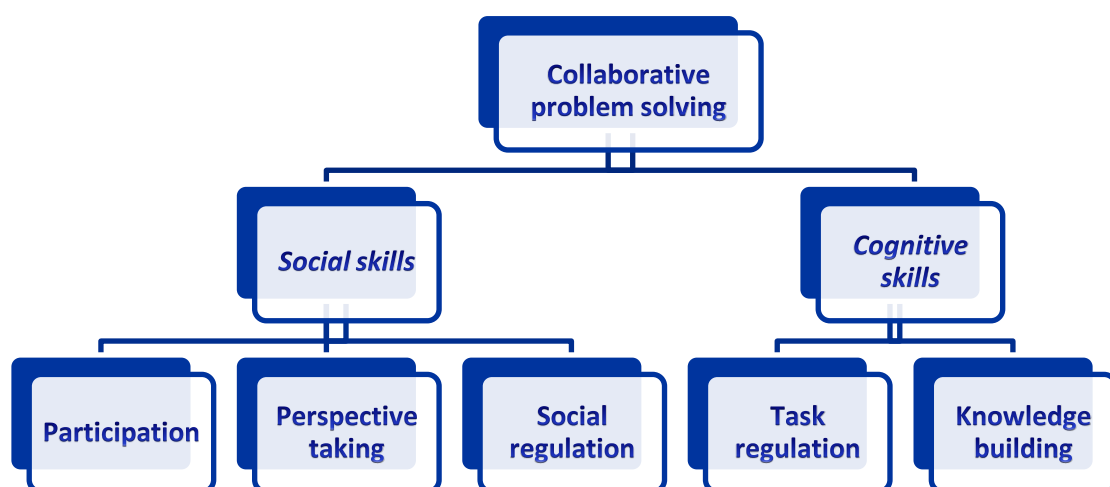


Figure 1. Collaborative problem solving construct (Griffin, Care, & McGav, 2012).

## **Gender, Grade and School Differences in CPS**

As research on CPS is limited, there lacks sufficient information about possible differences in CPS abilities according to the affecting factors of gender, grade and school. Some research has shown that different two-way interactions occur between individuals and their environments on a school-by-school basis; this is referred to as school engagement (Lekholm, 2011; Linnakylä & Malin, 2008). Studies have shown that parents' demographic characteristics have a strong influence on student achievement, as do school learning environments (Creemers & Kyriakides, 2010; Thapa, Cohen, Guffey, & Higgins-D'Alessandro 2013).

Differences between the genders have been observed to manifest in students' study performance, motivation or attitudes towards learning. Previous studies have noted that girls typically possess better academic skills and are more motivated towards learning (e.g. OECD, 2016). Zimmerman and Martinez-Pons (1990) examined student differences in self-regulated learning per several variables, including gender, based on interviews with fifth-, eighth- and eleventh-graders. They found that girls employed self-monitoring, goal setting, planning and study environment structuring more often than did boys. Chyung (2007) concluded that female students improved their self-efficacy more and scored significantly higher on final exams than did male students. Other studies assessed school culture and curricula; some found these factors to favour boys (Skelton, 2001), while others observed the opposite (Gentry, Gable, & Rizza, 2002). In this study, gender differences will be examined in terms of students' abilities in CPS.

Student abilities are assumed to rise with grade or year level when comparing students within the same school system. Grade level is usually equivalent to more years of education and better capacity due to developmental growth. Still, some research has shown that goal orientation and self-regulated learning skills do not always increase linearly with grade levels (Tang & Neber, 2008). Previous research has drawn a clear link between computer-supported collaboration and self-regulation (see Järvelä & Hadwin, 2013); the present study adds to this by examining the differences among students' aggregated CPS abilities across and within grade levels and schools to investigate possible school-based effects on student skills. The limitations of these comparisons will be addressed in the results section of this paper.

## **Aims and Research Questions**

While Finnish students identify valuing social skills and collaboration as their most important learned skills (Ahonen & Kinnunen, 2015), school leaders and teachers in comprehensive schools report that twenty-first century skills are not well-defined in their teaching methods and are difficult to teach (Niemi, 2012). Collaborative work is part of everyday instruction in Finnish schools, but there is no systematic intervention to actively teach and/or assess collaboration or CPS skills. Students tend to learn these skills more through natural maturation than specific teaching (Ahonen & Kankaanranta, 2015).

The CPS tasks were used to assess students' skill levels in the five strands indicated in Figure 1. The field trial using the ATC21S™ tool was organised to collect sufficient data to finalise the scoring and calibration of the assessment tool and thereby provide a picture of students' CPS skills. CPS tasks were piloted in four phases: concept checking, cognitive laboratories, classroom pilots and a trial. All tasks were translated and localised to suit the national purposes. The present study aims to answer the following two research questions in the context of Finnish schools:

1. Do items from online tasks measure Finnish students' skills across social and cognitive dimensions in a comparable manner to the total cohort involved in the project?
2. How do the CPS skills of participating students differ by gender and grade level within and between schools?

The scored data were used to analyse the differences in student ability by gender, grade and school. This type of analysis was performed for the Finnish student sample and provides original research on the specifics of differences measured using ATC21S™ CPS tasks. This kind of measurement of 'soft' skills is still in its early stages, therefore results should be interpreted with caution as tasks created in a different manner or based on a different foundational theory may lead to different results. All reported differences in student ability per gender, grade and school should be taken at face value only; there is no theoretical reason for differences to be generalisable to the wider population of students outside the five schools that participated voluntarily in the project. Limitations to the interpretation of the results in this study will be addressed at the end of results section.

## **Method**

### ***Participants***

A convenience sample of 480 students from Finland participated in the trial study between November 2011 and May 2012. The respondents were aged 11 to 15 years and were selected from five comprehensive schools representing school grades five (n = 107), seven (n = 157) and nine (n = 228). Gender distribution was even, with 51.8 percent of participants being male. School principals were contacted directly to discuss their interests and were offered the possibility to participate in the study on a voluntary basis. The scheduling and administration of the assessments were planned carefully with the school principals and involved teachers. In schools with up to ten participating groups of students, the trial study period took two weeks. All schools were visited during the trial study period, and a researcher was present at the assessment sessions.

The participating schools were heterogenic in enrolment and size. Two schools—school A, with 85 participating students (500 enrolled students overall) and school B, with 52 participating students (350 overall)—were lower secondary schools consisting of students in grades seven to nine. School C was a comprehensive school consisting of grades four to nine, with 182 participating students (330 overall). Two schools—school D, with 84 participating students

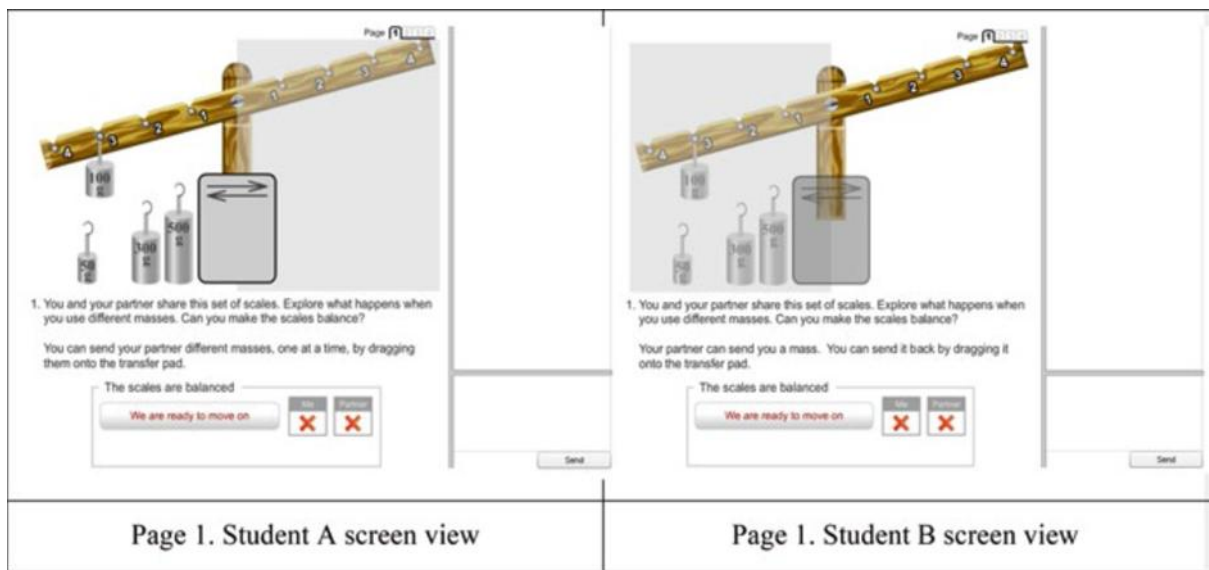
(300 overall) and school E, with 76 participating students (1,000 overall)—consisted of all grades from one to nine. Participating students were in grades five (schools B, C and E), seven (schools A, B, C and E) and nine (schools A, C, D and E).

To compare the skills measured in the Finnish sample with those from the other participating countries, data from the 480 Finnish students were compared with results published for the overall cohort of 4,078 students tested in the ATC21S™ project from six countries (including Finland).

### **Materials**

Students solved CPS tasks in pairs over the Internet during school hours. Each pair of students completed one bundle (three to four tasks). Each assessment session lasted 90 minutes. The assessment tasks were constructed to have the characteristics of problems that require true collaboration and relate to everyday teaching and learning (Care, Griffin, Scoular, Awwal, & Zoanetti, 2015). The tasks were designed for a student pair (e.g., students A and B), who were expected to communicate only through an online chat system. Each of the tasks began with an introduction, followed by a task one to eight pages in length.

The tasks were designed to maintain student engagement (Care et al., 2015). Many of the tasks were asymmetric, providing different information and resources to the respective students, to encourage collaboration. Symmetric tasks also required collaboration, as students had to determine whether they had the same information and answer a joint problem. Some tasks were content-dependent, while others were content-free. Content-dependent tasks did not necessarily require students to understand the content *a priori*, though they were framed in content disciplines such as mathematics or physics. Figure 2 illustrates an example content-dependent CPS task about physics. In this ‘beam balance’ task, the students had to balance the beam. Student A was given four weights and student B none. Two weights had to be passed to student B, at which point both students would place their respective weights in the correct notches to balance the beam. The students were assessed individually on their actions and collaboration via the chat box during the problem-solving process. Neither student could solve the task independently; the structure of the task created an ambiguous space for students to explore various methods of moving the weights.



**Figure 2. An example of a content-dependent CPS task (Beam Balance)**

This task example provides an excellent opportunity to explain how missing data was applied to the indicator coding to assess individual students' CPS abilities, even though the students were required to work together. To continue with the previous example task, if student A did not collaborate and never sent student B any useful weights, Student A would achieve a 0 score for the social indicator 'weights sent to partner on first page'. Student B might receive a 1 for the social indicator 'requests weights from partner on first page' even if student A never sent the weights. However, subsequent indicators for student B, such as the cognitive 'student B places weights on balance beam', would be labelled as 'missing data' rather than coded as 0 because student B never had the opportunity to place a weight on the balance beam. In traditional testing terms, student B 'never saw the question'. Per the Rasch analysis technique to calculate student ability, missing data was used to ensure that the assessment was fair and valid for both students. If one student depended on the other student to act, and this partner did not complete the necessary part of the task, the first student was never scored down in consequence. Validity arguments for the accuracy of the CPS tasks were not the focus of this paper, but this general mechanism for scoring was applied so that the gender and school differences discussed in the results could be interpreted with consideration for possible limitations in student scoring accuracies.

In the ATC21S™ portal, student completion of assessment tasks yielded log file data (see Figure 3). The generated data were captured in a process stream data file, and patterns in these data were coded automatically as indicators of CPS elements (Adams et al., 2015; Hesse et al., 2015). These tasks captured social and cognitive components of students' CPS skills. Each of the skills could thus be scaled based on the actions taken by the students. The actions were collected as process data alongside copies of the online chat discussions that took place between the students while performing CPS tasks.

"15:56:32",0,B,"Session join","IPAddress: 202.176.202.139; lang: en; playerName: S11_war01B; "	A so how should we do it;
"15:57:19",0,A,"Session join","IPAddress: 202.176.202.139; lang: en; playerName: s11_war01a; "	B put 1 at the top;
"15:58:02",1,B,"CameraConfirm","Cameras: {}; fullcoverage: false; ncams: 0; "	B can put at the side?;
"16:00:02",1,A,"CameraPlace","Posx: 3; Posy: 4; direction: S; fullcoverage: false; ncams: 1; "	A i think can;
"16:00:06",1,A,"CameraMove","Oldx: 3; Oldy: 4; Posx: 3; Posy: 2; direction: S; fullcoverage: false; ncams: 1; "	B now put at the side;
"16:00:12",1,A,"CameraMove","Oldx: 3; Oldy: 2; Posx: 3; Posy: 4; direction: S; fullcoverage: false; ncams: 1; "	A ok should we at the other side too?;
"16:00:13",1,A,"CameraRemove","Posx: 3; Posy: 4; direction: S; fullcoverage: false; ncams: 0; "	B the bottom;
"16:00:24",1,A,"CameraPlace","Posx: 6; Posy: 2; direction: W; fullcoverage: false; ncams: 1; "	B 1 whole line;
"16:01:09",1,A,"Send message","message: i think we should put the cameras at the side and the top; "	A what do you mean;

**Figure 3. An example log file from the Warehouse task.**

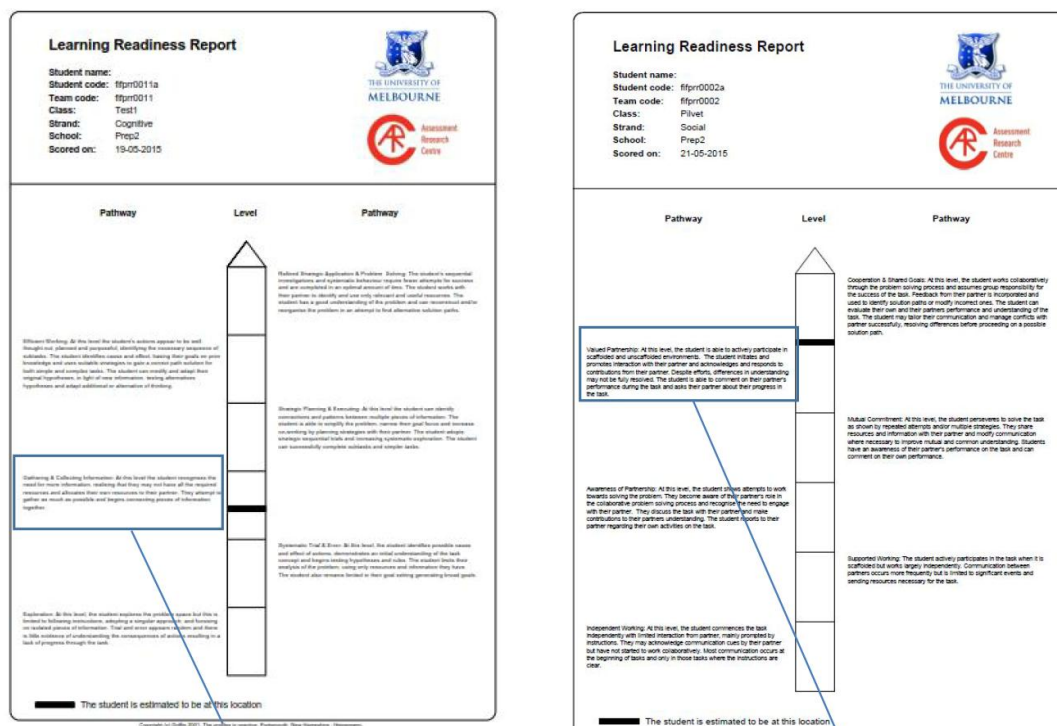
The tasks determined the CPS skills of students, and the observed indicators of CPS ability were designated *a priori* as social or cognitive components and as belonging to one of the five strands of subskills. The scoring itself took into consideration students' actions as they moved through the tasks. The process data consisted of distinct keystrokes and mouse events that indicated exploration of the task environment—such as typing, clicking, dragging, cursor movements, hovering time and action sequences—all of which explicitly demonstrated students' thinking processes and skill levels.

The log file data from the assessment tasks were processed at the Assessment Research Centre of the University of Melbourne. Student actions and chat messages were recorded and time-stamped. The files generated for the automatic records of student-task interactions were labelled as session log files (Adams et al., 2015). The subsequent log file data analysis used algorithms to turn the data into dichotomous or polytomous indicators based on the five-strand framework. The indicators were then analysed by exploring the Rasch model fit to the data (Harding & Griffin, 2016). Indicator fit to the model demonstrated that the construct could be interpreted at one, two and five dimensional levels to reflect the theoretical framework. That is, the indicators fit the multi-dimensional models (two or five), as well as the unidimensional model of CPS as a whole (Harding & Griffin, 2016).

The data was modelled after different dimensional theories so that student CPS skill progressions could be written for each dimension: cognitive, social and each of the five subskills. The analysis used in this study describes the interpretation



based on the two-dimensional model. A primary purpose of the ATC21S™ project was to identify student skill levels in societal and cognitive continua (Griffin, Care, & Harding, 2015). Learning readiness on the two-dimensional model was represented in both the social and cognitive categories. Learning readiness describes the level that student is placed on the scale according to his/her CPS abilities based on tasks completed ATC21S™ portal. The underpinning idea is to describe the level that she/he is at the present, and also point out what she/he is ready learn next. Learning readiness levels were presented as Rocket reports (see Figure 4), the lowest level being Level 1 and the highest Level 6.



**Figure 4.** Learning readiness reports from the ATC21S™ portal. The bold bar on the 'Rocket' graphic represents each student's level of readiness. Enlarged descriptions are included for Level 3 on the cognitive report and Level 5 on the social report.

## Analysis

Student CPS abilities were estimated based on 16,898 student task responses for 11 different tasks involving 4,078 secondary school students from Australia,

Costa Rica, Finland, the Netherlands, Singapore and the USA (co-author and collaborator, 2016). Item difficulty and student ability were calculated as marginal maximum likelihood estimates, obtained using an Expectation Maximisation, EM algorithm and applying plausible values when necessary (as described by Mislevy, 1991). These estimates were then reported on the same 'logit' scale, where both indicator difficulty and student ability were expressed in logits. Data was explored as a multi-dimensional construct consisting of both social and cognitive latent dimensions using the multi-dimensional, random coefficients multi-nominal logit model described by Adams, Wilson and Wang (1997). This process set the average task indicator difficulty to 0, while basing student ability estimates on the model. Variations between the difficulty of the indicators on the social and cognitive dimensions were illustrated using a variable map, which placed students (represented by X) and indicators (represented by item numbers) on a single logit scale (see Figure 5).

Student ability estimates were calculated based on their results on the indicators available to them per the bundle of tasks they completed. Analyses were performed using ConQuest™ (Wu, Adams, & Wilson, 2007), a multi-aspect test software that engaged a partial credit model with a Gauss-Hermite Quadrature estimation with 15 nodes (Bock & Lieberman, 1970). The students' abilities were calculated as Warm's weighted likelihood estimates (WLEs). Indicators were separated into social and cognitive constructs. The reliability of item and person separation (EAP and PV, respectively), reported for the sample of Finnish students, were obtained by dividing the variance of expected individual *a posteriori* ability estimates by the estimated total variance of latent abilities (Wu et al., 2007). SPSS v22 was used to investigate confidence intervals (using T-tests) and descriptive statistics. The statistical significance of the mean WLE differences among the student gender and grade groups between and within schools was tested using One Way ANOVA variance analysis and post-hoc tests. The correlations were counted using the non-parametric Spearman's rho. Statistical significances were represented by the p-value.

## Results

### *Student Abilities*

A variable map has been drawn to illustrate the interpretation of the construct at the hypothesised two-dimensional level (see Figure 5). A suitable set of indicative behaviours were scaled based on the social and cognitive dimensions of the Finnish students' CPS capacities. The chosen items were consistent in difficulty across the groups of students, the participating six countries and language versions (Harding & Griffin, 2016).

In the Finnish sample, person separation reliability (EAP-PV) was 0.73 in the social dimension and 0.72 in the cognitive dimension, comparable with the total cohort's person separations of 0.75 and 0.77, respectively, as reported by collaborators and co-author (2015). Item separation reliability was 0.98, indicating a spread of indicator difficulties that accounted for the Finnish student's abilities satisfactorily. The correlation between social and cognitive dimensions was 0.77, and 59% of the variance in student ability was shared

between the dimensions. Students who scored high in the cognitive dimension were more likely to obtain a high score in the social dimension and vice versa, indicating an association between the dimensions. Learning readiness levels were created empirically from the separate scaling of the dimensions using clusters of indicators which conceptually described increasing levels of skill proficiency. The scaling and creation of learning readiness levels were described previously by Griffin et al. (2015).

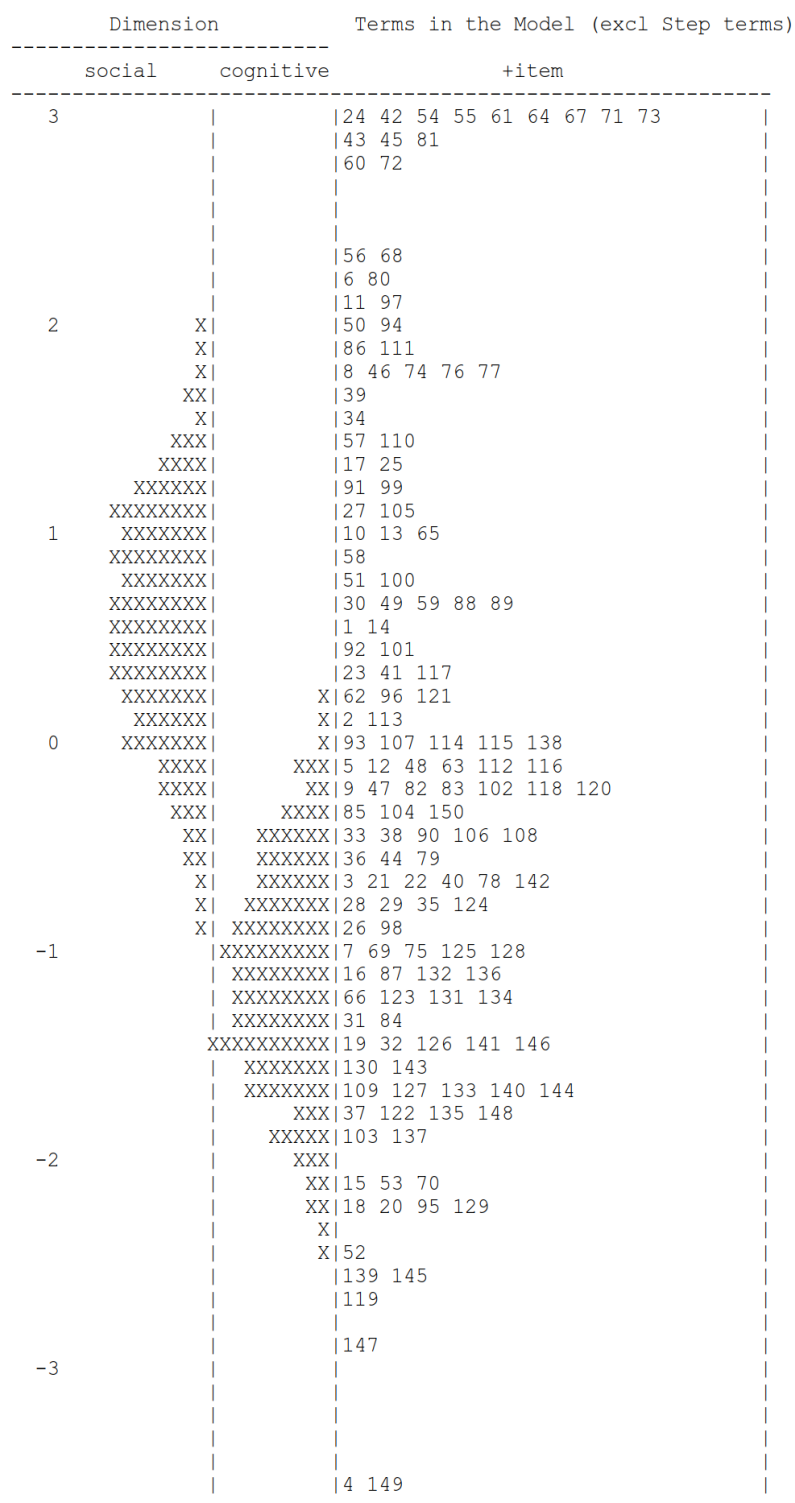
The student ability estimates (WLEs) required to achieve each learning readiness level are shown in Table 1. In the logit scale used to report student abilities, lower logit scores represented students with lower abilities. Indicator difficulties were calibrated on the same scale, such that a lower indicator difficulty (delta parameter) corresponded to an easier indicator. The mean difficulty of all indicators across both social and cognitive dimensions, were constrained to 0 logits, allowing student ability estimates to vary.

**Table 1. Range of student ability estimates (WLE) in the ATC21S portal corresponding to learning readiness levels.**

Level of Learning Readiness	Social	Mean	Cognitive	Mean
	WLE Range (logits)	Social WLE/SE	WLE Range (logits)	Cognitive WLE/SE
6	above 1.5		above 2.1	
5	0.3 to 1.5	0.537/ 0.029	1.7 to 2.1	
4	-0.5 to 0.3		0.5 to 1.7	
3	-0.7 to -0.5		-0.8 to 0.5	
2	-1.3 to -0.7		-3.5 to 0.8	-1.124/ 0.029
1	below -1.3		below -3.5	

*Note.* The mean ability for each dimension with a standard error is listed.

The aggregated mean of social latent ability for the tested population was 0.537 logits (SE = 0.029), representing Level 5 on the learning progression scale. The aggregated mean score for cognitive latent ability was -1.124 (SE = 0.029), or Level 2 on the learning progression scale (see Table 1). These results demonstrate that items connected with the cognitive dimension were more difficult for Finnish students than those connected with the social dimension (shown visually in Figure 5), which was also the case for the total cohort (Griffin et al., 2015). This does not necessarily mean that student abilities were higher in terms of social collaboration (compared to cognition); rather, the set of social indicators targeted low-level skills, while the cognitive indicators targeted high-level skills. This is accounted for in the written descriptions of learning readiness (see Figure 5) as empirically and conceptually determined levels of competence.



Each 'X' represents 3.9 cases

**Figure 5. Two-dimensional, single parameter variable map. Distribution of item difficulty is shown by numbered indicators on the right. Student ability distributions describing the social and cognitive dimensions of CPS are shown on the left, where each X represents 3.9 students.**

### *Aggregated gender and grade differences*

The results indicate that on an aggregate level, task difficulty did not vary among student grades or genders (see Figure 6). In the cognitive dimension, the mean of participating male students (N = 244) was -1.11 logits (SD = 0.84), and the female student mean was -1.11 logits (SD = 0.79). In the social dimension, the mean of participating male students was 0.52 logits (SD = 0.80), and the female student mean was 0.53 logits (SD = 0.76). There were no statistical differences in performance by gender for CPS based on the participants in this study.

Student ability estimates from each grade level were aggregated for all schools, and differences between grades were analysed. The fifth-grade students performed similarly to the seventh- and ninth-grade students in terms of cognitive and social CPS (see Figure 6). The comparisons run via One Way ANOVA and post-hoc tests indicated no significant differences in student CPS abilities by grade level. On aggregate level the fifth-grade students' mean latent ability (WLE) in the social dimension of CPS was 0.53 logits (SD = 0.70), the seventh-grade students' was 0.48 logits (SD = 0.69) and the ninth-grade students' was 0.49 logits (SD = 0.87). The mean latent ability in the social dimension for all grades fell within Level 5 on the learning readiness report. In the cognitive dimension, the mean logits were -1.11, -1.10 and -1.14 for grades five, seven and nine, respectively. The mean latent ability in the cognitive dimension for all grades fell within Level 2 on the learning readiness report.

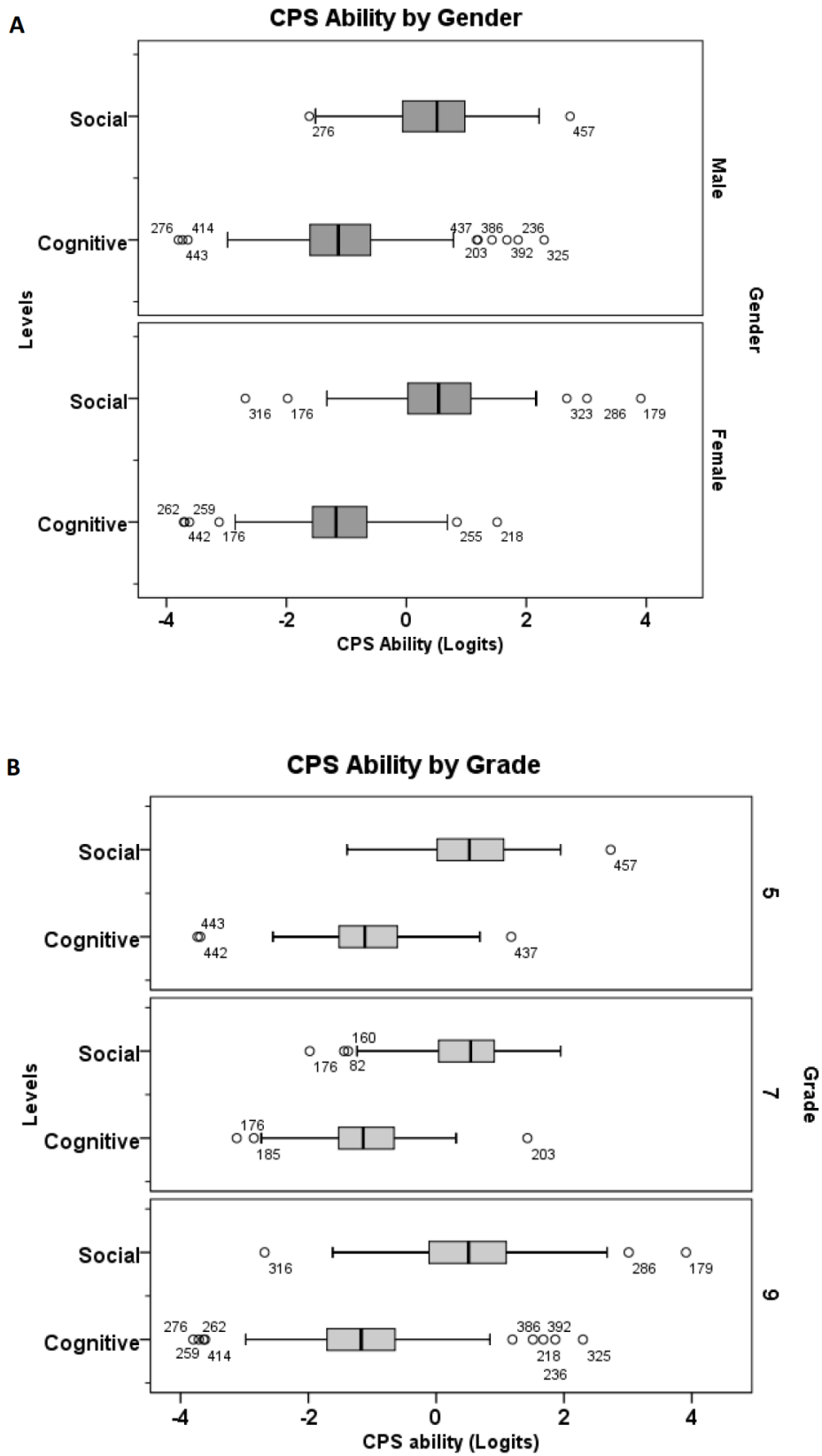


Figure 6. Students' CPS abilities by gender (A) and grade (B)

### *Grade differences across and within schools*

Mean student ability levels for the different grades were compared among and within the schools. Table 2 presents the number of students (n), the student ability estimate mean (WLE) and the standard deviation of the estimates for each grade in each school. Schools C and E had students participate from all three grades; school B had students participate from grades five and seven; and school A had students participate from grades seven and nine. In school D, only ninth-grade students were assessed (see Table 2).

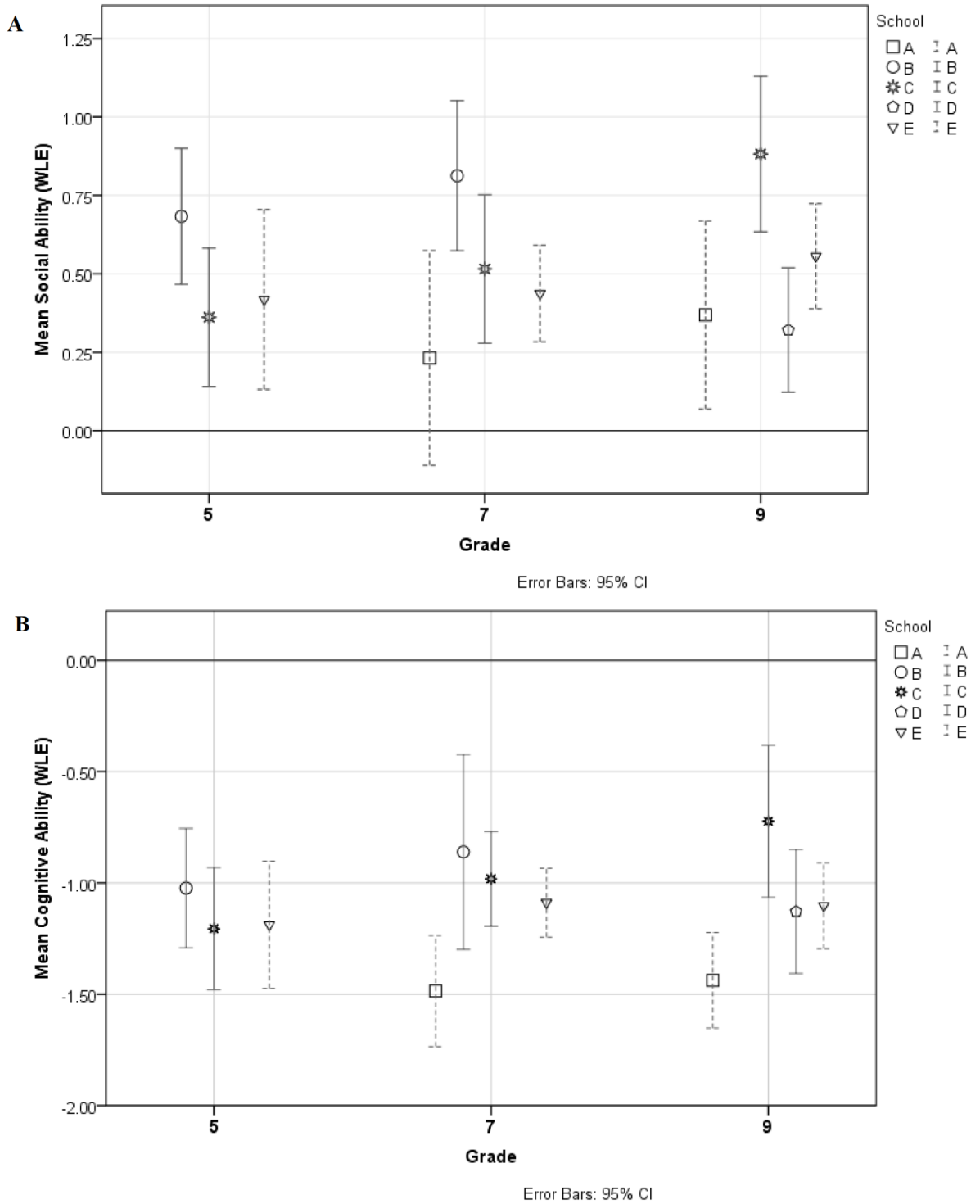
**Table 2. Student ability estimates across and within schools by grade.**

	Social skills			Cognitive skills		
	n/WLE means/SD			n/WLE means/SD		
	Grade 5	Grade 7	Grade 9	Grade 5	Grade 7	Grade 9
School A		27/ <b>0.23</b> /0.86	58/ <b>0.37</b> /1.14		27/ <b>-1.49</b> /0.63	58/ <b>-1.43</b> /0.82
School B	47/ <b>0.68</b> /0.74	24/ <b>0.81</b> /0.56		47/ <b>-1.02</b> /0.91	24/ <b>-0.86</b> /1.04	
School C	24/ <b>0.36</b> /0.52	30/ <b>0.52</b> /0.63	30/ <b>0.88</b> /0.66	24/ <b>-1.20</b> /0.65	30/ <b>-0.98</b> /0.57	30/ <b>-0.72</b> /0.91
School D			52/ <b>0.32</b> /0.71			52/ <b>-1.13</b> /1.00
School E	27/ <b>0.49</b> /0.72	72/ <b>0.44</b> /0.66	84/ <b>0.56</b> /0.77	27/ <b>-1.19</b> /0.72	72/ <b>-1.09</b> /0.66	84/ <b>-1.10</b> /0.89

The results indicate that the mean ability estimates varied among the grades within the schools (see Table 2 and Figures 7A and B), even though these differences were not apparent when comparing aggregate grade level data (Figure 6B). In each school, a higher grade resulted in a higher mean logit score within the school. Differences in mean student ability were just as great in a single grade across multiple schools as they were across multiple grades in a single school (see Figures 7A and B). For example, students from grade five in school B scored higher than the ninth-grade students from schools A, D and E in terms of social skills (see Figure 7A; Table 2). However, these differences were almost as great within the individual schools as they were among multiple schools. Statistically significant differences ( $p = 0.011$ ) between grades five and nine were observed in school C, where the ninth-grade students' average score in social skills was 0.88 logits (0.52 logits higher than the same school's fifth-grade students' average score of 0.36) (see Figure 7A). The ninth-grade students from school C were the only group to reach Level 3 for cognitive learning readiness; all the other groups' average scores were in Level 2 (see Table 1).

There was also a statistically significant ( $p = 0.003$ ) Spearman correlation between grade and social skills (WLEs) in school C. In terms of cognitive skills, there were neither statistically significant differences nor correlations within the individual schools. When comparing the mean differences among schools per grade level, some statistically significant differences were found. School A's seventh-grade students scored 0.58 ( $p = 0.029$ ) logits lower in social skills and 0.62 ( $p = 0.023$ ) lower in cognitive skills than the seventh-grade students in school B. When comparing ninth-grade students, those in school D scored 0.56

( $p = 0.046$ ) logits lower than students in school C. The cognitive skills of the ninth-graders in school A came in 0.71 ( $p = 0.007$ ) logits lower than those at school C.



**Figure 7. Students' mean CPS ability by grade per school. Mean social skills ability is represented by (A) and mean cognitive skills ability by (B).**



## Conclusions

Student performance of social and cognitive capabilities can be reported to teachers in levels on a learning progression so that teachers can provide interventions when appropriate. In the present study, student skills in the two dimensions of CPS were not affected by school, gender or grade at the whole data level. Upon examining the data from each school separately, differences were observed between grade levels within the individual schools. Higher-grade students performed better in CPS than lower-grade students. However, there were some unexpected differences of ability levels when skills per grade level were compared among the schools. School A, for example, reported lower levels of CPS skills than did the other participating schools at each grade level. The sample was opportunistic rather than probable, meaning the data did not allow any further analysis to determine possible reasons for the lower levels in that particular school.

The finding that skill levels increase by grade within a school is not a surprising result, as this tends to be the case for any content area. The differences in student achievement on the learning progression by grade support that the underlying construct is being measured accurately, as it is expected that as students are practicing these skills in school, their performance improves. The data suggests that some schools support their students in learning CPS skills more than others. Had a larger sample of schools been involved, aggregate differences in student performance in the two dimensions may have been observed by grade.

## Discussion

According to leaders and teachers in Finnish comprehensive schools, twenty-first century skills are not yet well established as teachable topics, are difficult to teach and do not have a clear role in school curricula (Kartovaara, 2009; Niemi, 2012). School system development requires better embedding of twenty-first century skills, as well as new assessment tools and teaching and learning methods, to equip citizens to function in a knowledge society (Krokkfors, Kangas, Vitikka, & Mylläri, 2010). Still, the results of recently published PISA 2015 Collaborative Problem solving assessment show that Finnish students were ranked on 7<sup>th</sup> place, with 534 average score points. That was not statistically significantly lower than Hong Kong (541 points) on 3<sup>rd</sup> place (OECD, 2017). Interesting finding of PISA study was also that Finland's gender difference was the greatest of all participating countries, when girls outperformed boys with 48 score points (OECD, 2017). The present study's finding with no difference between genders on Collaborative problem solving skills is very interesting and this indeed forms an interesting topic for future research.

One speculation could be that, based on researcher's observations, school A did not exhibit as high quality class orchestration during the data collection periods as did the other participating schools. Research has shown that orchestrating productive collaborative learning requires a balance between improvisation and organisation, as well as attention to the needs of different groups (Hämäläinen & Vähäsantanen, 2011). These speculative challenges to class orchestration may have had implications for the results at school A. Overall however, this study's

results indicate that differences in CPS skill levels are better compared on an individual school basis than across multiple schools. Factors subjective to each school, such as learning and teaching cultures, may have affected the way the students participated in the online tests and collaborated with each other.

The ATC21S™ portal assessment's outcomes provide teachers with information about student performance in terms of CPS skills. This marks the first time that this kind of information is available to teachers and the scientific public (Binkley et al., 2012; Csapo et al., 2012). The ATC21S project builds on the developmental model of learning (Vygotsky, 1978); thus, its primary goal is to maximise the developmental progression of individuals' skills, builds on learners' existing knowledge to reach higher and deeper levels of learning. However, in terms of formative assessment, this use of data-based evidence to improve learning and instruction has proved a challenge because there has been a general delay in delivering the analysed data to instructors for these purposes (Griffin et al., 2012).

### **Research limitations**

The present results should be interpreted with caution due to the study's sampling constraints; the selection of students to complete the tasks at each school was not controlled, and the schools themselves were not sampled appropriately for comparisons to be generalised across the Finnish population. The purpose of the current analysis, therefore, was to uncover patterns and possible differences that might be investigated further with appropriate sampling.

### **Suggestion for further research**

Recent developments in the field of large-scale assessment, namely PISA 2015 study, (OECD, 2017) have focused on measuring students' CPS skills. For example, psychometrically oriented assessment research aims to develop reliable methodologies to measure individuals' CPS skills on a large scale. To make these skills measurable requires splitting them into small sub-components and amalgamating their outcomes. The challenge in assessment in general, including assessing CPS, is to avoid oversimplifying the richness of collaboration. The CPS tasks and/or indicators may well be improved in the future as more is understood of the depth of the skills involved in CPS, and new techniques may be designed to capture and measure these skills.

The group orientation represented in this process-oriented research on computer-supported collaborative learning offers the beginnings of an understanding of the processes and contexts of productive collaboration. However, there is a need for stronger integration of and open communication between process-orientated and psychometrically orientated research communities (Woods, Mountain, & Griffin, 2015). The ways in which technology-enhanced formative assessment tools and systems might be used as pedagogical tools during collaborative processes constitute a significant area for future research.

## References

- Adams, R., Vista, A., Scoular, C., Awwal, N., Griffin, P., & Care, E. (2015). Framework for teachable collaborative problem solving skills. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 115–132). Dordrecht: Springer. [https://doi.org/10.1007/978-94-017-9395-7\\_2](https://doi.org/10.1007/978-94-017-9395-7_2)
- Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1–23.
- Ananiadou, K., & Claro, M. (2009). *21st century skills and competences for new millennium learners in OECD countries* (OECD Education Working Paper No. 41). Paris: OECD. <https://doi.org/10.1787/218525261154>
- Ahonen, A.K., & Kankaanranta, M. (2015). Introducing assessment tools for 21st century skills in Finland. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (Vol. 2). (pp. 213–226). Dordrecht: Springer. [https://doi.org/10.1007/978-94-017-9395-7\\_10](https://doi.org/10.1007/978-94-017-9395-7_10)
- Ahonen, A.K., & Kinnunen, P. (2015). How do students value the importance of twenty-first century skills? *Scandinavian Journal of Educational Research*, 59(4), 395–412. <https://doi.org/10.1080/00313831.2014.904423>
- Binkley M., Erstad, O., Herman J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Dordrecht: Springer. [https://doi.org/10.1007/978-94-007-2324-5\\_2](https://doi.org/10.1007/978-94-007-2324-5_2)
- Bock, R. D., & Lieberman, M. (1970). Fitting response model for dichotomously scored items. *Psychometrika*, 35, 179–197. <https://doi.org/10.1007/bf02291262>
- 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. 85–104). Dordrecht: Springer. [https://doi.org/10.1007/978-94-017-9395-7\\_4](https://doi.org/10.1007/978-94-017-9395-7_4)
- Chyung, S.Y. (2007). Age and gender differences in online behavior, self-efficacy, and academic performance. *The Quarterly Review of Distance Education* 8(3), 213–222.
- Csapo, P., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 143–230). New York, NY: Springer. [https://doi.org/10.1007/978-94-007-2324-5\\_4](https://doi.org/10.1007/978-94-007-2324-5_4)
- Creemers, B., & Kyriakides, L. (2010). School factors explaining achievement on cognitive and affective outcomes: Establishing a dynamic model of educational effectiveness. *Scandinavian Journal of Educational Research*, 54(3), 263–294. <https://doi.org/10.1080/00313831003764529>
- Gentry, M., Gable, R., & Rizza, M. G. (2002). Students' perceptions of classroom activities. Are there grade-level and gender differences? *Journal of Educational Psychology*, 94(3), 539–544. <https://doi.org/10.1037//0022-0663.94.3.539>
- Gordon, J., Halász, G., Krawczyk, M., Leney, T., Michel, A., Pepper, D., . . . Wiśniewski, J. (2009). *Key competences in Europe: Opening doors for lifelong learners across the school curriculum and teacher education* (CASE Network Report No. 87). Warsaw: CASE Centre for Social and Economic Research. <https://doi.org/10.2139/ssrn.1517804>
- 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* (pp. 1–16). Dordrecht: Springer. [https://doi.org/10.1007/978-94-007-2324-5\\_1](https://doi.org/10.1007/978-94-007-2324-5_1)
- Griffin, P., Care, E., & Harding, S.M. (2015). Task characteristics and calibration. In E. Care, P. Griffin, & M. Wilson (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 133–177). Dordrecht: Springer. [https://doi.org/10.1007/978-94-017-9395-7\\_7](https://doi.org/10.1007/978-94-017-9395-7_7)

- Greiff, S., Niepel, C., & Wistenberg, S. (2015). 21<sup>st</sup> century skills: International advancements and recent developments. *Thinking Skills and Creativity*, 18, 1–3. <https://doi.org/10.1016/j.tsc.2015.04.007>
- Harding, S.M.E., Griffin, P.E., Awwal, N., Alom, B.M.M., Scoular, C. (2017). Measuring collaborative problem solving in a mathematics context, *AERA Open*. <https://doi.org/10.1177/2332858417728046>
- Harding, S.M.E., & Griffin, P.E. (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: Springer. [https://doi.org/10.1007/978-94-017-9395-7\\_2](https://doi.org/10.1007/978-94-017-9395-7_2)
- Hämäläinen, R., & Vähäsantanen, K. (2011) Theoretical and pedagogical perspectives on orchestrating creativity and collaborative learning. *Educational Research Review*, 6, 169–184. <https://doi.org/10.1016/j.edurev.2011.08.001>
- Järvelä, S., & Hadwin, A. (2013). New frontiers: Regulating learning in CSCL. *Educational Psychologist*, 48(1), 25–39. <https://dx.doi.org/10.1016/j.chb.2015.01.050>
- Kartovaara, E. (2009). *Opetuksen järjestäjien ja rehtoreiden näkemyksiä ja kokemuksia perusopetuksen vuoden 2004 opetussuunnitelmauudistuksesta* [Experiences from curriculum change in 2004]. Helsinki: Opetushallitus.
- Krokkfors, L., Kangas, M., Vitikka, E., & Mylläri, J. (2010). Näkökulmia tulevaisuuden koulupedagogiikkaan [Insights into future school pedagogy]. In R. Smeds, L. Krokkfors, H. Ruokamo, & A. Staffans (Eds.), *InnoSchool – välittävä koulu. Oppimisen verkostot, ympäristöt ja pedagogiikka* [InnoSchool – caring school. Learning networks, learning environments and pedagogy] (pp. 51–85). Espoo: Aalto-yliopisto.
- Lekholm, A. K. (2011). Effects of school characteristics on grades in compulsory school. *Scandinavian Journal of Educational Research*, 55(6), 587–608. <https://doi.org/10.1080/00313831.2011.555923>
- Miller, M., & Hadwin, A. (2015) Scripting and awareness tools for regulating collaborative learning: Changing the landscape of support in CSCL. *Computers in Human Behavior*, 52, 573–588. <https://dx.doi.org/10.1016/j.chb.2015.01.050>
- Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56, 177–196. <https://doi.org/10.1007/bf02294457>
- Niemi, E. K. (2012). Opettajakysely. In E. K. Niemi (Ed.), *Aihokokonaisuuksien tavoitteiden toteutumisen seuranta-arviointi 2010* [Evaluation of the fulfilment of cross-curricular themes 2010] Koulutuksen seurantaraportit 2012:1 (pp. 19–39). Helsinki: Opetushallitus.
- OECD. (2016). *PISA 2015 results (Vol. 1): Excellence and equity in education*. Paris: OECD Publishing. <http://dx.doi.org/10.1787/9789264266490-en>
- OECD. (2017). *PISA 2015 Results (Volume V): Collaborative Problem Solving, PISA*, OECD Publishing, Paris. <http://dx.doi.org/10.1787/9789264285521-en>
- OECD. (2017). *PISA 2015 collaborative problem solving framework*. Retrieved from <http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf>
- Pepper, D. (2011). Assessing key competences across the curriculum and Europe. *European Journal of Education*, 46(3), 335–353. <https://doi.org/10.1111/j.1465-3435.2011.01484.x>
- Rotherham, A. J., & Willingham, D. (2009). 21st century skills: The challenges ahead. *Educational Leadership*, 67(1), 16–21.

- Silva, E. (2009). Measuring skills for 21st-century learning. *Phi Delta Kappan*, 90(9), 630–634. <https://doi.org/10.1177/003172170909000905>
- Skelton, C. (2001). *Schooling the boys: Masculinities and primary education*. Buckingham: Open University Press.
- Strijbos, J.-W. (2011). Assessment of (computer-supported) collaborative learning. *IEEE Transactions on Learning Technologies*, 4(1), 59–73 <https://doi.org/10.1109/tlt.2010.37>
- Tang, M., & Neber, H. (2008). Motivation and self-regulated science learning in high-achieving students: Differences related to nation, gender, and grade-level. *High Ability Studies*, 19(2), 103–116. <https://doi.org/10.1080/13598130802503959>
- Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A review of school climate research. *Review of Educational Research*, 83(3), 357–385. <http://dx.doi.org/10.3102/0034654313483907>
- Vygotsky, L. S. (1978). *Mind and society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.
- Woods, K., Mountain, R., & Griffin P. (2015). Linking developmental progression to teaching. In P. Griffin, & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 267–292). New York, NY: Springer. [https://doi.org/10.1007/978-94-017-9395-7\\_14](https://doi.org/10.1007/978-94-017-9395-7_14)
- Wu, M. L., Adams, R. J., & Wilson, M. R. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software*. Melbourne: ACER Press.
- Zimmerman, B. J., & Martinez-Pons, M. (1992). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology* 82(1), 51–59. <https://doi.org/10.1037//0022-0663.82.1.51>