Validation of the PISA 2015 collaborative problem-solving competence measure

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Humanities

2022

Sofia Eleftheriadou

School of Environment, Education and Development

Contents

Final word count: 75,420 words

List of Tables

List of Figures

List of Abbreviations

AARE	Australian Association for Research in Education
ACER	Australian Council for Educational Research
BERA	British Educational Research Association
CI	Cognitive Interviewing
CPS	Collaborative Problem Solving
DIF	Differential Item Functioning
EERA	European Educational Research Association
ESCS	Economic, Social, and Cultural Status
ESRC	Economic and Social Research Council
GCSE	General Certificate of Secondary Education
IRT	Item Response Theory
MNSQ	Mean-Square
OECD	Organisation for Economic Co-operation and Development
PISA	Programme for International Student Assessment
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RQ	Research question
SEED	School of Environment, Education and Development
TIMSS	Trends in International Mathematics and Science Study

Abstract

Collaborative problem solving (CPS) as a competence has received much attention in the educational literature, especially after the release of the Programme for International Student Assessment (PISA) 2015 results. In PISA 2015, 15-year-olds' competence to work in collaborative settings was assessed across countries. The validity of the PISA 2015 CPS competence measure has been repeatedly questioned, mainly due to the constraints imposed in the computer-based assessment. This thesis critically examines the validity of the CPS competence assessment as instrumentalised in the PISA 2015 study by analysing student responses and reflections on the PISA 2015 CPS items.

A mixed methods approach was mobilised by linking analysis and results from research phases that use quantitative and qualitative methodologies. This thesis draws on the unified validity framework of Messick (1989) as well as the literature in CPS competence in education to investigate validity and seek interpretation of test scores. Two systematic literature reviews focused on the conceptualisation and operationalisation of CPS and the assessment of CPS competence. The first empirical phase of the thesis involved the use of the Rasch measurement framework to analyse the PISA 2015 dataset for England (a secondary data analysis). Analysing the available secondary data, largely unused to date, this phase examined a validation based on the multidimensional character of CPS competence. As a next step, the constructed CPS competence measures were used as variables in further statistical analyses to evaluate external and consequential aspects of validity. The second empirical phase of the thesis involved primary data collection through cognitive interviews and verbal probing with students from a secondary school in England. Using the released PISA 2015 CPS assessment task in new ways (cognitive interviewing), this phase adds to what PISA/OECD have already published/reported.

Results suggest that: a) the identification of student response processes revealed limitations to the validity of the CPS task items used, b) the associations of CPS with theoretically relevant variables did not provide sufficient evidence to support the external and structural validity aspects of the CPS competence measures, and c) several weaknesses were identified in the instrument, the PISA methodology and reporting, which eventually undermined its external and consequential validity. The thesis concludes that data derived from the PISA 2015 CPS competence assessment should be treated with caution, suggesting that test score interpretation should recognise that the assessment only reflects student CPS competence when working with computer-simulated partners in a restricted assessment environment. The study's implications highlight the importance of considering evaluation in real-life situations and provide insight into how the use of such instruments in high stakes testing environments might contribute to the implementation of standardised curricula. Overall, the present study stands as an independent validation of the PISA 2015 CPS competence assessment, identifying threats to validity that weaken extrapolation from the CPS competence assessment to real-world collaboration situations.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

Dedication

To the loving memory of my grandmother Evanthia who passed away from COVID-19 on 29th November 2021. I couldn’t say a proper goodbye to you, but this is to thank you for being the strong woman I got to know, practising feminism in your life, and believing in the power of education.

Acknowledgements

There are several people who supported and encouraged me towards the completion of this study, and I am grateful to them.

First, I would like to express my gratitude to my first supervisor Maria Pampaka for her support and guidance throughout the years of my doctoral study. Maria, I cannot thank you enough for your commitment with my work and your encouragement when I found it hard to keep going. Your mentoring throughout this process gave me the space to develop as a researcher and persist with the completion of this thesis. Thank you very much for the opportunities to get involved in research projects and your support to do so many things during my doctoral studies. Similarly, I would like to thank Julian Williams for being intellectually challenging, for his comments in my work, and for helping me think more critically. Thank you, Julian, for always being so approachable and supporting my internship in Melbourne, Australia. I would also like to thank Alexandru Cernat for his comments in my work and his advice in developing my thesis.

I would like to thank the Australian Council for Educational Research (ACER) for hosting my internship in Melbourne, Australia and for providing a welcoming space to discuss my research. Special thanks go to Ray Peck and Claire Scoular for helping me organise the internship and for making my experience in Melbourne so valuable. My thanks also go to the rest of the researchers at ACER, who were very generous with their time and happy to have chats about my PhD over a nice flat white. This internship was made possible by the financial support from the Economic and Social Research Council (ESRC) - North West Social Science Doctoral Training Partnership (NWSSDTP), and I feel grateful for the opportunity I was given.

I would also like to thank the teacher who gave me access to their school and the students who participated in my study (as well as their parents/guardians for giving permission to do so).

I would like to thank colleagues at the Manchester Institute of Education, University of Manchester, particularly Laura Black and Jenna Mittelmeier, for supporting my research training during my doctoral studies, Carlo Raffo and Pauline Prevett, for giving me opportunities to be involved in their research projects, and Clelia Cascella for the conversations we had around Rasch measurement that helped me with my data analysis. Studying at MIE, I was lucky to share my PhD journey with amazing PGRs, so I would like to thank them for being genuinely nice and supportive. Artemis Christinaki and Choen Yin Chan (Helen), I cannot thank you enough for being the best PhD buddies I could have and for your support to finish this thesis.

I thank Kyriakos Neanidis for helping me navigate the PhD application process and for his advice on my professional development, Ioanna Nedou for reminding me the small things that are important in life, Natasa Kazakou, Kiki Stepani, and Foteini Thoma for being such supportive friends all these years. I thank Michalis Constantinides with all my heart for being by my side and believing in finishing this thesis. Finally, I thank my family for their support and care as I have been pursuing my studies.

This work was supported financially by the Economic and Social Research Council (ESRC) [NWSSDTP Grant Number ES/P000665/1]. This research would not have been made possible without this financial support, so I am grateful for the opportunity I was given. I also thank The University of Manchester for awarding me the prestigious President's Doctoral Scholar Award in recognition of my commitment and desire to work on a challenging research project.

The Author

Sofia completed her undergraduate degree BSc (Hons) in Education Studies at the University of the Aegean, Greece in 2014. She then moved to Manchester, UK to pursue her studies at postgraduate level. She first completed a MA in Inclusive Education and Special Educational Needs at the Manchester Metropolitan University in 2015, followed by a MSc in Social Research Methods and Statistics at the University of Manchester in 2017. After graduating with a distinction, Sofia continued to the PhD Education at the University of Manchester and was awarded on 2017 two prestigious scholarships: (i) the +3 Studentship from the Economic and Social Research Council (ESRC) and (ii) the President's Doctoral Scholar Award, the University of Manchester. Both awards are given to outstanding research students, who demonstrate academic excellence and leadership potential, to support research methods training and research.

During her studies, Sofia has worked as a researcher in various research projects detailed below. Currently, she is working as a Research Associate at the University of Manchester in the project 'Socioeconomic disadvantage and the attainment gap' funded by the Education Endowment Foundation. In this role, she conducts a systematic literature review on the relationship between socioeconomic disadvantage and the attainment gap in the English education system. Through her involvement in research projects, Sofia has gathered a wide range of experience and developed her expertise in conducting systematic literature reviews and measure validation. Her research experience is summed up next, followed by a list of awards, academic journal publications (including the four papers presented in the thesis), conference presentations, and other research-related activity.

Research experience

Research Associate - Manchester Institute of Education, The University of Manchester, UK

- Project title: 'Socioeconomic disadvantage and the attainment gap: Review of the relationship between socioeconomic disadvantage and the attainment gap in the English education system' (funded by Education Endowment Foundation, 2022)
- Project title: 'Unpacking and measuring financial literacy self-efficacy of millennial mature students in the UK' (funded by the University of Manchester, 2022)

Research Assistant - Manchester Institute of Education, The University of Manchester, UK

- Project title: 'Mathematics Education 5-14 and Gender: Review of affect, attitudes, and aspirations', (funded by Joint Mathematical Council, 2021)
- Project title: 'Increasing Competence and Confidence in Algebra and Multiplicative Structures (ICCAMS)', (funded by Education Endowment Foundation, 2018)
- Project title: 'Unsettling Understandings of Maths Anxiety, Systematic literature review', (funded by British Academy, 2017)

Awards

- Best Paper Award 2019 Emerging Researchers' Conference, European Educational Research Association (EERA)
- Presentation Award 2020 Emerging Career Researcher (shortlisted), British Educational Research Association (BERA)

Professional qualifications and fellowships

- Associate Fellow (AFHEA), Advance Higher Education (UK), 2021
- Qualified Teacher Status (QTS), National College for Teaching and Leadership, 2014

Peer-reviewed journal articles

Eleftheriadou, S. (in preparation, co-authoring with supervisory team). Conceptualisation of collaborative problem solving: a systematic literature review.

Eleftheriadou, S. (in preparation, co-authoring with supervisory team). Assessment of students' collaborative problem-solving competence with the use of computer-simulated, scenario-based tasks: a systematic literature review.

Eleftheriadou, S. (in preparation, co-authoring with supervisory team) Dimensionality and validity of the PISA 2015 collaborative problem-solving competence construct.

Eleftheriadou, S. (in preparation, co-authoring with supervisory team) Validity of PISA 2015 collaborative problem-solving assessment based on student response processes.

Scoular, C., **Eleftheriadou, S.,** Ramalingam, D., & Cloney, D. (2020). Comparative analysis of student performance in collaborative problem solving: What does it tell us? *Australian Journal of Education*, 64(3), p. 282-303. doi:10.1177/0004944120957390

Conference contributions

Eleftheriadou, S. (2022). Assessment of Students' Collaborative Problem-solving Competence: A Systematic Literature Review. European Conference for Educational Research (ECER). Yerevan State University, Armenia, 22-25 August.

Eleftheriadou, S. (2021). Examining construct representation and dimensionality of the PISA 2015 collaborative problem-solving measure. European Conference for Educational Research (ECER). Online, 9 September.

Eleftheriadou, S., & Pampaka, M. (2020) The Relationship Between Student Attitudes and Collaborative Problem Solving: Evidence from PISA 2015 in England [Paper Session]. American Educational Research Association (AERA) Annual Meeting San Francisco, CA (Conference Cancelled).

Eleftheriadou, S., & Pampaka, M. (2020). Examining construct representation and dimensionality of the PISA 2015 collaborative problem-solving measure. European Conference for Educational Research (ECER) (Conference cancelled).

Eleftheriadou, S., & Pampaka, M. (2020). Measuring collaborative problem solving internationally and its association with student learning outcomes. British Educational Research Association (BERA) (Conference cancelled).

Eleftheriadou, S., & Pampaka, M. (2019). Examining evidence for the validity of PISA 2015 collaborative problem-solving measure using the Rasch model. Australian Association for Research in Education (AARE). Brisbane, Australia. 1-5 December.

Eleftheriadou, S. (2019). Conceptualisation and measurement of collaborative problem solving. British Educational Research Association (BERA). Manchester, UK. 10-12 September.

Eleftheriadou, S. (2019). Conceptualisation and measurement of Collaborative problem solving: a systematic review of the literature. Emerging Researchers' Conference (ERC) European Educational Research Association (EERA). Hamburg, Germany. 2-3 September.

Eleftheriadou, S. (2019). Exploring the measurement properties of the PISA 2015 collaborative problem-solving measure. Annual UK Rasch User Group Meeting. Cambridge, UK. 21 March.

Eleftheriadou, S. (2019). An exploration of student responses to the PISA 2015 collaborative problem-solving assessment: a mixed-methods approach. Methods X Conference. Liverpool, UK. 17 May.

Eleftheriadou, S. (2019). An exploration of student responses to the PISA 2015 collaborative problem-solving assessment: a mixed-methods approach. SEED PGR Conference 2019, The University of Manchester, 21 May.

Cascella, C., Lei, K. H., Pampaka, M., **Eleftheriadou, S.,** & Williams, J. (2018). Math anxiety around the world. British Society for Research into Learning Mathematics (BSRLM). London, UK. 10 November.

Pampaka, M., **Eleftheriadou, S.,** Cascella, C., Estevez, S. and Lei, K. H. (2018) Mathematics Anxiety around the globe. British Academy conference 'Contemporary research in mathematics anxiety and emotions: advancing the field'. London, UK. 9 March.

Pampaka, M., & **Eleftheriadou, S.** (2017) Measuring emotions and teaching practices and the association between the two. British Educational Research Association (BERA). Brighton, UK. 5-7 September.

Other research-related activities

- Extracurricular university service (Co-convenor of Emerging Researchers' Group, European Educational Research Association – EERA, 2022)
- Internship (Australian Council for Educational Research - ACER, Melbourne, 2019)
- Reviewer in academic journals (Research in Mathematics Education; Mind, Culture, and Activity: An International Journal)
- Reviewer in conferences (American Educational Research Association - AERA Annual Meeting, European Conference for Educational Research - ECER)
- Conference organised (SEED PGR Conference 2018, Role: Organising committee member, The University of Manchester, UK, 22 May)
- Seminars organised ("Evaluation, Measurement and Assessment Interdisciplinary Seminars series", funded by Collaborative Innovation Grant, Methods North West, 2019)

Chapter 1 Introduction: Motivation and purpose of thesis

1.1 Introduction

Problem solving has been assessed for several decades, often following Polya's (1945) problem-solving process which consisted of four steps: understanding the problem, devising a plan, carrying out the plan, and looking back. The topic of collaborative problem solving (CPS) competence has recently attracted interest in international and national assessments of student performance (Care et al., 2018; Graesser et al., 2018; OECD, 2017a). There are several reasons explaining the growing interest of researchers in the CPS concept. Due to global changes in the workforce, being able to solve problems individually is no longer seen as sufficient (Autor et al., 2003). Many businesses and industries around the globe place high value on collaboration, problem solving, and interpersonal skills in their employees to deal with the demands of technological advances and a globalised workforce (Rios et al., 2020). As the job market demands CPS skills, education systems in different parts of the world have progressively begun to incorporate these into their curricula and teaching approaches (Care, Anderson, et al., 2016; Creese et al., 2016), and to promote problem-based and collaborative learning as a means for more effective student learning (Hmelo-Silver et al., 2013). Educational research has shown the benefits of engaging in collaborative activities (Andrews-Todd & Forsyth, 2020; Dillenbourg & Traum, 2006; Gillies, 2016). Furthermore, international and national assessments have been found to focus on measuring and developing student CPS skills (Fiore et al., 2017; Griffin et al., 2012; National Research Council, 2011; OECD, 2017a).

In 2015, a large-scale international assessment of CPS competence was conducted by the Organisation for Economic Co-Operation and Development (OECD), when the CPS domain was introduced in the Programme for International Student Assessment (PISA) study. Driven by the perceived needs of policy, the PISA study aimed to measure, and consequently ensure that students are equipped with, skills to meet the CPS demands of their future careers (OECD, 2017a). For the purposes of PISA 2015, CPS competence was defined as "the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution" (OECD, 2017a, p. 134). This definition is important throughout the thesis, especially in its interpretations of the validity of testing/measurement. Testing "capacity" is particularly difficult in general but coming to 'a solution' is particularly problematic in this 'group' context.

The definition of CPS competence, for the purposes of the PISA 2015 CPS assessment, incorporates three core competencies unique to PISA's CPS: Establishing and maintaining shared understanding, Taking appropriate action to solve the problem, and Establishing and maintaining team organisation (OECD, 2017a). "Establishing and maintaining shared understanding" relates to keeping track of what other team members know about the problem, their perspectives, and a shared vision of the problem states and activities. "Taking appropriate action to solve the problem" relates to performing actions, which can include physical actions and communication acts, that follow the appropriate steps to achieve a solution. "Establishing and maintaining team organisation" relates to helping to (re)organise the group by considering the knowledge, skills, abilities, and resources of group members, following the rules of engagement for roles in the group, as well as handling obstacles (OECD, 2017a). These three newly conceptualised competencies are crossed with the four individual problem-solving processes (i.e., exploring and understanding, representing and formulating, planning and executing, and monitoring and reflecting) to create a matrix of 12 cells, each representing a specific CPS skill (Table 1.1). The individual problem-solving processes have been previously defined in the PISA 2012 framework following Polya's (1945) work related to problem solving in the context of mathematics.

Following the framework illustrated in Table 1.1, PISA assessed students' CPS skills using an individualised computer-simulated assessment focusing on students' performance in a collaborative event, as opposed to group performance. Each item included in the CPS assessment is classified as targeting one of the CPS skills, and thus it can be mapped back to one of the three core competencies. The main rationale for the PISA 2015 CPS assessment was the need for a standardised summative assessment system, designed to provide large-scale information to countries about their student populations' achievements. By assessing 15-year-olds' CPS competence, PISA aimed to address the lack of internationally comparable data in this field, and provide policy makers with information that would assist them to develop programmes to improve students' CPS skills (OECD, 2017b).

Table 1.1. PISA 2015 Collaborative problem-solving framework (OECD, 2017a)

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

Meanwhile, many research groups around the globe focussed on how to assess students' CPS competence, creating computer-simulated, scenario-based tasks, and investigating methods of data analysis appropriate for interpreting students' social interactions (e.g., von Davier et al., 2017). Published computer-based assessments of CPS competence include designs in which students navigate through tasks as prescribed by multiple-choice pathway options from computer-simulated partners (OECD, 2017b; Rosen & Foltz, 2014). Other assessments include tasks in which students use free form chat to interact with other students (Griffin et al., 2015; Scoular & Care, 2019). The timeliness and increasing interest in the assessment of CPS competence, is further highlighted by the recent publication of special issues in journals, such as Applied Measurement in Education (Greiff & Kyllonen, 2016b), Journal of Educational Measurement (von Davier, 2017), and Computers in Human Behavior (Graesser et al., 2020). Although the number of assessments aimed to measure students' CPS competence has increased over time, a systematic overview/review of the available literature has been lacking to date. There is therefore a need for a comprehensive synthesis of these assessments to inform policy and practice about the state-of-the-art in the field.

In CPS assessment tasks, the group member knowledge, gender, personality, motivation, and other demographic characteristics have been listed as likely significant factors in group composition (Webb & Gibson, 2015). In the PISA CPS assessment, computer-simulated partners were programmed to represent team members with different roles, attitudes, levels of competence, as well as behaviour (e.g., team members supporting and praising others versus team members interrupting and negatively criticising the work of others) thus varying the situations students are confronted with (OECD, 2017a). This approach has been argued to allow for a high degree of control and standardisation required for measurement, and it has been adopted by other studies as well (e.g., Hsieh & O'Neil, 2002; O'Neil, 1999; Rosen & Foltz, 2014). However, whether computer-simulated partners can be designed to reliably mimic realistic conversational partners or the extent to which interacting with computer-simulated partners generalises to interacting with human partners remains an open question (Webb & Gibson, 2015).

Several authors recognised limitations in the constrained task designs employed for the assessment of students' CPS competence including the fact that they deviate from real-life collaboration environments (Graesser et al., 2018; Scoular et al., 2017). It has been argued that it is difficult to develop problem-solving tasks using computer-simulated partners that appear genuine and natural for the students without first trialling such tasks with real students (Scoular et al., 2017). Questions regarding the extent to which computer-simulated partners could fully capture the real collaboration between humans remain unanswered (Rosen, 2015; Scoular & Care, 2020). The year 2022 marks 110 years since the birth of Alan Turing, a mathematician and computer scientist, whose work has been highly influential in the field of artificial intelligence[1]. In his ground-breaking paper "Computer Machinery and Intelligence", Turing introduced his hypothetical test, what is now known as the Turing test, which aimed to answer the question 'Can machines think?' (1950, p. 433). His method was used to determine whether a machine is capable of exhibiting human-like behaviour. Miller (as quoted in Tallentire & Shervin, 2022), points out that artificial intelligence-powered robots are not yet able to develop such deep learning and level of complexity, compared to humans, hence are not able to respond to unknown or unseen scenarios with the same level of logic, experience, or reasoning. Whether artificial lifeforms can learn to feel, have empathy, and develop 'human-like' intuition and instinctive behaviour remains an open question (Miller as quoted in Tallentire & Shervin, 2022).

Since the release of PISA 2015 CPS results, hardly any studies have made use of the secondary data related to students' CPS competence or provided validity evidence for the test score interpretation and use (i.e., De Boeck & Scalise, 2019; Scoular, Eleftheriadou, et al., 2020; Tang et al., 2021). Specifically, Tang et al. (2021) used the PISA 2015 data for four provinces in China to explore the factors predicting students' CPS competence. Scoular, Eleftheriadou, et al. (2020) compared three different computer-based CPS assessments (including PISA 2015) and the derived

[1] https://www.mub.eps.manchester.ac.uk/science-engineering/2022/06/23/passing-as-human-what-is-the-turing-test/

student CPS competence measures to investigate the extent that they measure the same construct. The data they used were from Australian student samples and they found that skills related to negotiation and audience awareness were not well represented across the three assessments. DeBoeck and Scalise (2019) used the PISA 2015 data for the United States to investigate the relationship between CPS performance and the invested time and number of actions in collaborative episodes. They found that students showing a fast trial-and-error strategy on the assessment were not very successful compared to those choosing a slower, more thoughtful response style. However, since PISA has not provided any interpretive information from the student response process (such as think-aloud protocols), it was not possible to examine why students were doing what they were doing (De Boeck & Scalise, 2019).

Given that the publishing of the PISA 2015 CPS results will likely increase the attention received from researchers, educators, and policy makers on students' CPS competence, validity evidence related to the adequacy and appropriateness of interpretations and actions based on PISA 2015 CPS competence scores is needed urgently. Following this, the thesis attempts to address this lack of knowledge by systematically exploring the validity of the PISA 2015 CPS competence measure using student samples and mixed data analyses from England (as will be further justified next).

1.2 Personal motivation for this research

I first came across the PISA study when I was an undergraduate student in Greece studying for a BSc in Education. In a lecture of the Science Education module, one of the famous "league" tables of PISA results comparing students' performance in science among countries was presented. After moving to the UK for postgraduate studies, I came across the PISA study and its results for the second time. It was in a statistical analysis module taught by Dr Alexandru Cernat (supervisory team) for the MSc Social Research Methods and Statistics, where I used PISA student scores to investigate relationships with other variables to learn various statistical analysis methods. In the same year, I also attended a seminar in Rasch measurement taught by Prof. Maria Pampaka and Prof. Julian Williams (supervisory team). Before attending that seminar, I was only using the PISA scores as they were provided by the survey organisers, often without even looking at question examples from which the scores were derived from. After the Rasch measurement seminar, I started thinking about scale construction, validity, and the ethics/consequences of measurement.

Specifically, I was thinking about the fact that it is not good enough to create a scale that will have good psychometric properties alone, but it is also important to think about how this scale will be used, making sure it will not have adverse consequences for students. Similarly, thinking about what could happen if a scale constructed gets out of control and results from an assessment are used for purposes other than the ones initially designed for. It was a fascinating seminar that introduced me to the sub-field of educational measurement and made me start thinking critically about the concept of validity. By that time, the OECD released a draft framework for its new assessment in the innovative domain of ‘collaborative problem solving’. I was surprised to find out that the assessment of such skills was conducted via a constrained individualised computer-based assessment, in which participating students worked with computer-simulated partners instead of real humans. Furthermore, communication was limited to the exchange of a selection of pre-defined messages. This design format raised immediately concerns about validity to me, especially since it was employed for the purposes of assessing something that is supposed to be a collaborative construct. This is how I became interested in the topic and why I was motivated to conduct this research at the doctoral level.

When developing the research proposal, PISA 2015 data for CPS competence had not been released yet. Initially, the research design was solely focused on the secondary data analysis of the PISA 2015 data, and specifically conducting cross-country comparisons. This was motivated by the work that I had completed for my MSc dissertation: “Maths anxiety, teaching practices and maths performance: Evidence from multilevel analysis of PISA 2012”. However, after the publication of

PISA 2015 results for CPS, in November 2017, it was made clear that item information was only available for a limited number of CPS items (12 out of 117 items included in the assessment). Comparability analysis was not possible to go in depth without this information, and therefore, I decided that my research design should focus in one country (England). The timing of PISA 2015 CPS data release coincided with the start of my PhD journey, so I had the opportunity to adapt the proposed research design. After being able to see the only one released PISA CPS assessment task (12 items), I was intrigued to research it more substantially, and this is how I ended up considering qualitative methods of inquiry and more specifically cognitive interviewing. England was chosen as a country for qualitative data collection due to the common language and access to secondary schools that made conducting interviews with students possible for me.

1.3 Purposes of the Thesis

The key purpose of this thesis is to investigate the validity of the PISA 2015 CPS assessment and the derived CPS competence measures. To explore this complex issue, four research questions (RQ) guided this research. As this thesis makes use of the concept of "collaborative problem solving", the first research question is more of a conceptual problem:

> RQ1. How has "collaborative problem solving" been conceptualised and operationalised in the educational research community?

To address RQ1, this thesis systematically explores different conceptualisations in the educational research literature in Chapter 4. Then, I ask:

> RQ2. How has students' CPS competence been assessed using computer-simulated, scenario-based assessment tasks in educational research studies?

Findings related to the conceptualisations of CPS from Chapter 4 are used in Chapter 5 to explain the view of CPS taken by researchers using computer-simulated, scenario-based assessment tasks to measure students CPS competence.

As mentioned previously, the number of assessments aimed to measure students' CPS competence has increased. To answer RQ2, the assessments of students' CPS competence are systematically reviewed in Chapter 5 to develop a comprehensive description of the state-of-the-art in the field and a critical evaluation of the validity evidence reported. Then, I pose the following research question:

> RQ3. What are the strengths and limitations of measurement validity of the CPS competence measure for England based on PISA 2015 data?

In relation to RQ3, different aspects of validity have been investigated in the literature to date in relation to the CPS competence measures derived from computer-simulated, scenario-based assessment tasks, some of which not adequately covered. In Chapter 6, validity evidence for the PISA 2015 CPS competence measure is investigated using secondary data from 15-year-old students in England taking the PISA 2015 study. A particular focus is placed on structural and external validity aspects, by investigating the hypothesised multidimensional structure of the CPS competence construct, as well as its relation to other supposedly relevant (collaboration and performance) constructs. Finally, I ask:

> RQ4. What does the PISA 2015 CPS assessment actually measure according to student perspectives?

Chapter 7 explores the evidence for the substantive validity aspect of this CPS competence measure. To address RQ4, it examines in depth what a small sample of secondary school students in England say about how they comprehend and explain their answers to the items included in one PISA 2015 CPS assessment task through cognitive interviewing. This method, although highly relevant for the validation of CPS competence measures, has been surprisingly neglected in the CPS-related literature so far. This study is the first of its kind (as far as I know from my literature review) in systematically exploring the validity of the PISA 2015 CPS competence assessment using cognitive interviews.

1.4 Structure of the Thesis

This thesis consists of a total of eight chapters. The Introduction chapter, as presented here, sets the background/gaps in the literature, and outlines the details of each scientific paper contained within the thesis. It also presents the rationale for following a journal format.

Chapter 2 offers contextual information about the educational system in England and discusses the influential role of the PISA study in policy. The chapter also presents the theoretical background regarding the investigation of validity over the years as well as current validity issues and criticisms concerning the PISA study and more specifically the PISA 2015 CPS assessment.

Chapter 3 presents the methodological approaches underpinning this thesis in which different (quantitative and qualitative) methods are applied on different data. The chapter details the methods used to conduct systematic literature reviews and analyse secondary (quantitative) and primary (qualitative) data, as well as the rationale behind such choices.

Chapters 4, 5, 6, and 7 are written as scientific papers, each tackling a research question (and several sub-questions detailed in 1.6 below with a summary of the purpose of each Research paper). This format allows for sections of the doctoral thesis to be written as research papers in a way that are suitable for publication in peer-reviewed journals. Specifically, Chapters 4 and 5 present results of two systematic literature reviews, and Chapters 6 and 7 present results of the two empirical phases.

Chapter 8 brings the thesis together by providing a critical evaluation of the findings and an overview of the contribution to knowledge that each chapter makes. It sets out how the field has moved forward and re-contextualises limitations. Finally, it highlights implications for future research, policy, and practice, before concluding with ideas for future work.

1.5 Rationale for a Journal Format Thesis

The thesis follows the guiding principles for a journal format thesis provided by the University of Manchester (Appendix 1). This format allows for sections of the doctoral thesis to be written as research papers in a way that is suitable for publication in peer-reviewed journals. In this thesis, four results chapters (Chapters 4, 5, 6, and 7) are written in the form of four self-contained scientific papers designed to stand alone. Each of these includes the following sections: abstract, introduction, background of relevant literature, methods, results, discussion, conclusion, and references. This format has enabled me to address with greater accuracy the key contributions made by each paper and present these as such to the educational research community. In addition, it offers quick dissemination of results. Finally, presenting research in the form of papers has helped to develop my skills in writing scientific papers, which is essential for a research/academic career.

1.6 Summary of the purpose of each Chapter/Research paper

In this section, an overview of each research paper in Chapters 4, 5, 6, and 7 is provided. Following the University of Manchester journal format thesis guidelines, this section includes the details of each paper contained within the thesis and explains how these papers constitute a coherent body of work and relate to each other.

1.6.1 Overview of Chapter 4: "Conceptualisation of collaborative problem solving: a systematic literature review"

Chapter 4 focuses on research question one (RQ1: How has "collaborative problem solving" been conceptualised and operationalised in the educational research community?), examining one sub-question:

> RQ1.a: How are the variations in the CPS conceptualisations explained by diverse research purposes?

Chapter 4 is concerned with systematically examining how CPS has been conceptualised in recent empirical and theoretical educational research. As a starting point, I use a body of literature examining classic theoretical approaches to the study of problem solving and collaboration, to ground CPS within the wider literature of concepts with substantial research history. Then, I argue that the existing literature reviews on the topic of CPS do not yet provide an in-depth insight into the variety of conceptualisations of CPS in the field and across educational settings. Thus, the intention of this chapter/paper is to obtain an overview of CPS conceptualisations within educational research to have a holistic picture of how the concept has been used. This is done for the purpose of informing future research (including meta-analyses and reviews) and, policy and practice, about the current state-of-the-art in the field. To accomplish such scope, a systematic literature review is used, where CPS definitions, theoretical underpinnings, research purposes, and methods for data collection and analysis, are gathered and analysed. To get a broader understanding of the literature, I was interested in finding out how researchers defined the construct, what theories they used to support their definitions and how they operationalised it in the empirical part of their work (where appropriate).

In this chapter/paper, I follow the steps proposed by Gough et al. (2016) and Petticrew and Roberts (2006) to guide my systematic literature review. Specifically, after developing the research questions that guide the review, I detail the article search approach that I follow to identify potential articles for the review, I present and justify the inclusion criteria that I apply when screening articles, and I detail the approach that I follow to extract and synthesise data, after concluding with the selection process. Using the article as the unit of analysis, I first apply descriptive codes to articles to enable me to map the size and nature of the literature, before moving to a more in-depth examination. Drawing from Thomas and Harden (2008), I use thematic synthesis to analyse the conceptualisation and operationalisation of CPS, as a "concept" in research practice, adopted in the articles reviewed. Constant comparison between articles is central in this approach, leading to the identification of three categories of conceptualisations: i) CPS competence in which CPS is defined as a competence that is needed for one to be an effective contributor to a CPS activity, ii) CPS practice in which CPS is seen as a type of pedagogic approach/intervention, and iii) CPS interaction in which CPS is conceived as activity taking place in a joint problem space. The dangers in privileging only research evidence using one category of conceptualisations is discussed, with a view to taking this work further in Chapter 5, as detailed next.

1.6.2 Overview of Chapter 5: "Assessment of students' collaborative problem-solving competence with the use of computer-simulated, scenario-based tasks: A systematic literature review"

Chapter 5 focuses on research question two (RQ2: How has students' CPS competence been assessed using computer-simulated, scenario-based assessment tasks in educational research studies?), examining three sub-questions:

> RQ2.a: What are the existing assessments of students' CPS competence and their characteristics (e.g., subject domain, task design features)?
> RQ2.b: Which facets of CPS competence do the assessments measure?
> RQ2.c: What strategies for validating CPS competence measures are reported?

Questions have been raised about the validity and authenticity of CPS competence assessments using computer-simulated, scenario-based tasks, while a systematic review targeting this topic is currently missing. Using Chapter 4 as a backdrop, Chapter 5 reviews existing assessments of CPS competence and the relevant validity evidence provided. It takes a closer look at the methods of data collection (i.e., assessment instruments) and analysis of articles assessing students' CPS competence with the use of computer-simulated, scenario-based tasks. This is done for the purpose of developing a better understanding and a critique of existing assessments of CPS competence. Following a systematic literature review methodology (Gough et al., 2016; Petticrew & Roberts, 2006) to address transparency and replicability, this chapter/paper contributes to knowledge about the assessment of students' CPS competence by: i) describing the characteristics of the existing CPS assessments, ii) categorising the facets of CPS competence targeted for measurement, and iii) evaluating the strategies adopted for validating the CPS competence measures.

To answer RQ2.a, I use the assessment as the unit of analysis and extracted information from the articles regarding the task design features (e.g., communication mode, scoring approach). To answer RQ2.b, I draw mainly from a framework of CPS competence proposed by Oliveri et al. (2017) to extract information about the facets (or skills) within components of CPS competence targeted by each assessment (unit of analysis). To answer RQ2.c, I draw from Messick's (1989) unified validity definition emphasising content, substantive, structural, generalisability, external, and consequential aspects for the analysis of validity evidence provided by each article (as the unit of analysis) in the review. The limitations in current assessments of CPS competence are discussed, and the gaps in the validity evidence concerning CPS competence measures are highlighted, with a view to taking this work further in Chapters 6 and 7.

1.6.3 Overview of Chapter 6: "Dimensionality and validity of the PISA 2015 collaborative problem-solving competence construct"

Chapter 6 focuses on research question three (RQ3: What are the strengths and limitations of measurement validity of the CPS competence measure for England based on PISA 2015 data?), examining three sub-questions:

> RQ3.a: To what extent is a hypothetical three-dimensional structure of 'establishing and maintaining shared understanding', 'taking appropriate action to solve the problem', and 'establishing and maintaining team organisation' measures supported empirically by the PISA 2015 data for CPS assessment in England?
>
> RQ3.b: To what extent are the constructed measures of CPS competence invariant across gender?

> RQ3.c: How are the constructed measures of CPS competence related to other relevant (collaboration and performance) constructs?

As a result of the criticisms around the validity of the PISA 2015 CPS assessment, presented in detail in Chapters 4 and 5, Chapter 6 examines the dimensionality and subsequently aspects of validity of the CPS competence measures. Most of the previous educational assessment research about the validity of students' CPS competence measures, which derived from computer-based assessments such as the PISA 2015 CPS assessment, is based on internal validity or reliability investigations, whereas aspects such as structural, external, and consequential validity are less investigated. At the same time, the validity of the PISA 2015 CPS competence measure has been repeatedly questioned, mainly due to the constraints in communication imposed in the assessment (e.g., Graesser et al., 2018; Scoular et al., 2017). In this chapter/paper, I draw from Messick's (1989) definition of validity as a unified concept and the Rasch measurement framework (Rasch, 1960), for the analysis of student responses to the newly developed items of PISA 2015 CPS assessment.

I examine evidence concerning aspects of validity of the CPS competence measures, based on the multidimensional character of CPS competence and using data from the PISA 2015 study targeting 15-year-olds in England. Drawing from the PISA 2015 CPS framework (OECD, 2017a), the three core competencies for CPS (i.e., Establishing and maintaining shared understanding, Taking appropriate action to solve the problem, and Establishing and maintaining team organisation) are used to investigate whether there is evidence suggesting that the CPS competence construct, as defined and assessed by the PISA 2015 CPS assessment, should be multi-dimensional. In the analysis, I examine construct validation by using a series of Rasch measurement models to construct and validate an overall CPS competence measure as well as sub-scales of CPS competence. As a next step, I use the constructed CPS competence measures as variables in further statistical analyses, including correlations and regression modelling to evaluate external and consequential aspects of validity.

Reflections about scale modifications, policy and practice implications and suggestions for future research are discussed. This chapter/paper sheds some light into the validity of the PISA 2015 CPS competence measure by using the available secondary data with a view to take this work further in Chapter 7, which explores the validity of the PISA 2015 CPS competence assessment using more in-depth qualitative methods, as described next.

1.6.4 Overview of Chapter 7: "Validity of the PISA 2015 collaborative problem-solving assessment based on student response processes"

Chapter 7 focuses on research question four (RQ4: What does the PISA 2015 CPS assessment actually measure according to student perspectives?), examining two sub-questions:

> RQ4.a: How do students comprehend the CPS assessment items and how do they explain their answers to them?
> RQ4.b: What are the implications for the external validity of the CPS assessment?

Chapter 7 examines student response processes and offers critiques of the external validity of the CPS competence assessment. Despite the substantial research reporting on the assessment of students' CPS competence, this is the first study, to the best of my knowledge, that explores the validity of the PISA 2015 CPS competence measure through cognitive interviews. Cognitive interviewing is employed as the main approach to collect qualitative data in Chapter 7 and as a method it examines the response processes and interpretations of respondents when answering survey questions (Willis, 2005). Adopting a definition of validity as a unified concept (Messick, 1995), I use elements of the grounded theory coding approach (Charmaz, 2006) for the analysis of student response processes to the newly developed, published items of PISA 2015 CPS assessment.

This chapter/paper also builds on the quantitative results derived from the analysis of PISA 2015 data for England (presented in Chapter 6). The focus is on what a small sample of secondary school students in England say about how they comprehend the items included in one PISA 2015 CPS assessment task and how they explain their answers to them. I draw on cognitive interviews using verbal probing with students interacting with the CPS competence construct.

OECD released the content from only one CPS task (out of six CPS tasks) included in the PISA 2015 CPS assessment. Because of this restriction, this chapter/paper makes use of the only publicly available PISA 2015 CPS task. Primary data are analysed in new ways that add to what PISA/OECD have published/reported so far. The findings point to several weaknesses in the instrument, the PISA methodology and reporting, and implications for its external and consequential validity.

1.7 References and appendices

Chapter 9 provides the full list of references that have been cited throughout the whole thesis. In accordance with the University of Manchester guiding principles for journal format thesis (Appendix 1), specific reference lists have been compiled for each Chapter, written as a scientific journal (Chapters 4, 5, 6, and 7), and are presented at the end of each chapter respectively. For that reason, there is some duplication across the various reference lists. Chapter 10 is the final chapter of the thesis and presents the Appendices.

1.8 Authorship credit

I have written this thesis with the support of my supervisors. Therefore, the four research papers (Chapters 4, 5, 6, and 7) are co-authored according to the level of contribution of each supervisor and are at different stages of peer review. Following the University of Manchester guiding principles for journal format thesis, the level of contribution of co-authors in each paper is made explicitly clear in this section. In general, I have taken the **major role in all aspects** of production of the four papers including planning and execution, data acquisition, data analysis, and writing.

In relation to **Chapter 4** (Research paper 1: CPS concepts review), I formulated the overarching research aims, defined research questions, developed a systematic literature review methodology, conducted the systematic literature search in literature databases, screened and selected articles for review, analysed data, and wrote the draft. Nevertheless, Prof. Maria Pampaka and Prof. Julian Williams contributed more extensively to the final product. Their involvement was mainly in defining theoretical gaps and helping to develop a systematic literature review methodology, as well as reviewing a final version of the draft and suggesting modifications. Prof. Maria Pampaka has also helped with screening randomly chosen articles to ensure that there is agreement in the way the inclusion criteria are used. This is the reason why they are both co-authors in this article.

In relation to **Chapter 5** (Research paper 2: CPS measurement review), I formulated the overarching research aims, defined research questions, developed a systematic literature review methodology, conducted the systematic literature search in literature databases, screened and selected articles for review, analysed data, and wrote the draft. Nevertheless, Prof. Maria Pampaka has helped with screening randomly chosen articles to ensure that there is agreement in the way the inclusion criteria are used. In addition, Prof. Maria Pampaka, Prof. Julian Williams, and Dr Alexandru Cernat contributed by editing and structuring initial versions of this paper, reviewing final version of the draft, and suggesting modifications. This is the reason why they are all co-authors in this article.

In relation to **Chapter 6** (Research paper 3: Rasch analysis of PISA 2015 CPS measure and its correlates with important variables), I formulated the overarching research aims, defined research questions, analysed data, and wrote the draft. Nevertheless, Dr Maria Pampaka helped with aspects of Rasch analysis and the interpretation of the results. I also received help by Prof. Maria Pampaka, Prof. Julian Williams, and Dr Alexandru Cernat in structuring the presentation and discussion of the data. They all contributed with editing and structuring initial versions of this paper, as well as by reviewing the final version of the draft and suggesting modifications. This is the reason why they are all co-authors in this article.

In relation to **Chapter 7** (Research paper 4: Cognitive interviewing of students), I formulated the overarching research aims, defined research questions, collected, and analysed data, and wrote the draft. Nevertheless, Prof. Maria Pampaka and Prof. Julian Williams contributed more extensively to the final product. Their involvement was mainly in defining theoretical gaps and helping to develop an interview protocol as well as analyse the student interview data. They have also contributed by reviewing the final version of the draft and suggesting modifications. This is the reason why they are both co-authors in this article.

1.9 Dissemination strategy and Publication plan

The plan for the research design and methodology adopted in this thesis, and described in more detail in **Chapter 3**, was presented at the Methods X Conference (2019) at the University of Liverpool (Eleftheriadou, 2019b).

I presented an early version of the paper in **Chapter 4** (Research paper 1: CPS concepts review) focusing on initial results from systematic literature review, at the British Educational Research Association (BERA) Annual Conference (2019) at the University of Manchester (Eleftheriadou, 2019c) as part of the symposium with the theme: 'What worked (or not): Synthesising evidence in the interplay of theory policy practice and research agendas.' I also presented preliminary results of the three categories of CPS conceptualisations at European Educational Research Association's (EERA) Emerging Researchers' Conference (2019) at Universität Hamburg (Eleftheriadou, 2019d). Following the Emerging Researchers' Conference, presenters are invited to hand in full papers for the Best Paper Award competition, which undergo a double-blind peer review process by a committee. I participated in the competition in 2019, and I was delighted that my paper titled "Conceptualisation and measurement of collaborative problem solving: a systematic review of the literature" was awarded the Best Paper Award 2019 (https://eera-ecer.de/about-eera/promoting-emerging-researchers/erc-best-paper-award/best-paper-award-2019/). After receiving this award, I was invited to write a blog post for the North West Social Science Doctoral Training Partnership providing an overview of the awarded paper (https://nwssdtp.ac.uk/2020/09/08/best-paper-award/). This paper is currently in preparation for submission to a peer-reviewed journal.

I presented preliminary results of the paper in **Chapter 5** (Research paper 2: CPS measurement review) at an organisation wide seminar at the Australian Council for Educational Research (ACER), Melbourne, Australia (17th October 2019), where I was hosted for an ESRC-funded internship. This paper is currently in preparation for submission to a peer-reviewed journal and has been accepted to be presented at the upcoming European Conference for Educational Research 2022 in Yerevan State University, Armenia (22-25 August). It is also worth noting that, due to my specialised interest in computer-simulated assessments of students' CPS competence and in particular, the PISA 2015 CPS assessment, I was invited by researchers at ACER to co-author a paper for the special issue of the Australian Journal of Education in the topic of: '20 Years of PISA in Australia: What can we say?'. In the paper titled: 'Comparative analysis of student performance in collaborative problem solving: What does it tell us?' published in 2020, I contributed by describing the assessment approach followed by PISA 2015 CPS assessment and by comparing it to two other assessments of student CPS competence (Scoular, Eleftheriadou, et al., 2020). It is important to note that the aforementioned paper does not form part of the thesis, in the sense that it is not one of the stand-alone research papers, since I was not the first (or single) author in that work. However, this paper is included in the selected articles for review in Chapter 5.

The paper in **Chapter 6** (Research paper 3: Rasch analysis of PISA 2015 CPS measure and its correlates with important variables), which is related to a secondary data analysis of PISA 2015 CPS assessment data at the student level, using the framework of Rasch measurement models, was presented in various versions at the Annual UK Rasch User Group Meeting (2019) at Cambridge Assessment, at the

Australian Association for Research in Education (AARE) Conference (2019) at Queensland University of Technology (Eleftheriadou & Pampaka, 2019), and at the European Conference for Educational Research (2021) held online (Eleftheriadou, 2021). This paper is currently in preparation for submission to a peer-reviewed journal.

The paper in **Chapter 7** (Research paper 4: Cognitive interviewing of students), which examines student cognitive interviews when responding to a PISA 2015 CPS assessment task, was presented at the School of Environment, Education and Development (SEED) PGR Conference (2019) at the University of Manchester (Eleftheriadou, 2019a). This paper is currently in preparation for submission to a peer-reviewed journal.

Chapter 2 The English context and the PISA study

2.1 Summary of the chapter

The objective of this chapter is to present the theoretical background and context regarding the investigation of validity. The influential role of the Programme for International Student Assessment (PISA) study in educational policy is discussed, followed by information about the educational English context including a brief overview of the school system and assessment practice in England. Issues related to the validity of PISA results and criticisms concerning the PISA (2015) study raised both in the UK and internationally are also discussed. The purpose is to clarify the 'modern' conceptualisation of validity of assessments such as PISA, and particularly to situate these in the context of their policy/political consequences.

2.2 The influential role of the PISA study in policy

International surveys of student performance such as PISA have gained increasing popularity around the world, as they intend to measure the performance of an education system (Baird et al., 2011). Since 2000, the Organisation for Economic Cooperation and Development (OECD) has been conducting the PISA study, a three-year cycle of 'curriculum-independent' standardised tests of reading, mathematics, and science literacy. PISA takes place at the end of compulsory schooling and has the overall goal of measuring what 15-year-old students can do with the knowledge they have acquired in school. Rather than attempting to assess pupils' knowledge of national curricula, PISA attempts to capture how well young people can apply reading, science, and mathematics skills in real-world situations. Additionally, in most PISA cycles, cognitive assessments of an "innovative domain" have been included, which are basically assessments of additional cross-curricular competencies (OECD, 2017a).

The innovative domain for the PISA 2015 study was collaborative problem solving (CPS), whereas for the PISA 2012 study, this was individual problem solving. Additionally, for the first time in PISA 2015, the main mode of assessment was computer-based tests. Apart from the cognitive assessment, questionnaires are also used in the PISA study to gather information from students and teachers about their background, attitudes, and teaching-learning environment. This information is used to contextualise the student attainment findings (Baird et al., 2011).

It has been argued that PISA has become a widely-watched indicator of national educational performance across the globe (Jerrim, 2021). Results from the PISA study have been used in education policy development in many countries (Baird et al., 2016; Hopfenbeck et al., 2018). Since 2000, the number of participating countries in PISA has increased from 32 countries and 200,000 students to 72 countries and 540,000 students in PISA 2015 (OECD, 2017a). In addition, PISA has been described as the OECD's platform for policy construction at a national, international, and possibly global level (Rizvi & Lingard, 2006). This international dimension of the survey, gives PISA a particularly significant weight as an indicator of the success or failure of education policy (Grek, 2009). There are several major players in PISA, other than the OECD itself. Surveys are coordinated by the governments of participating countries, assessment materials are developed by leading subject experts, and the fieldwork is designed and managed by an international study centre (Baird et al., 2011).

PISA is now considered as one of the most influential studies in education around the world, with its results having a substantial impact upon education policy as well as massive media coverage (Baird et al., 2011; Jerrim, 2021). Results from every PISA cycle are widely anticipated by academics, journalists, and policy makers (Hopfenbeck, 2016; Jerrim, Parker, et al., 2018). PISA results have previously led to reforms of education systems, including, for example, reforms to curriculum in Norway, South Korea and Mexico, along with changes to national assessments in Slovakia and Japan (Baird et al., 2011; Breakspear, 2012). In the United Kingdom, PISA has had an impact upon education discussion and debates (Baird et al., 2011; Jerrim, 2021). The National report of PISA 2015 results included intra-UK comparisons of educational performance, titled 'PISA across the UK' (Jerrim & Shure, 2016). Results comparing England, Northern Ireland, Scotland, and Wales in PISA 2015 were also reported within the national media (e.g., Adams et al., 2016). It has been argued that, due to a lack of accessible and comparable national examination data, relatively few comparisons have been conducted for student educational achievement in the United Kingdom (Jerrim & Shure, 2016). Therefore, PISA has become the “go-to” resource for intra-UK comparisons of students’ academic achievement (Jerrim, 2021; Jerrim & Shure, 2016). Given PISA’s prominent role in comparing and understanding educational performance across the UK, it has been argued that, it is vital to provide sound and reliable evidence upon which comparisons are made (Jerrim, 2021).

Although the impact of PISA may differ at the national level, potential mechanisms behind its ability to influence policymaking have been identified in the literature (Hopfenbeck et al., 2018). Specifically, several authors suggest that OECD’s status as an international organisation allows for powerful yet indirect governance through PISA (Bieber & Martens, 2011; Hopfenbeck et al., 2018). In addition, the OECD’s promotion of PISA data usage has been argued to endorse data-driven policymaking, encouraging systems to rely upon external authorities for knowledge production and policy guidance (Grek, 2010; Hopfenbeck et al., 2018; Meyer, 2014). Overall, the power and impact of PISA on educational policy appears to reinforce the need to further monitor and examine issues that carry weighty implications for consequential validity (Messick, 1989) not only at the system level but also at the individual level (Hopfenbeck et al., 2018).

2.3 Educational system and assessment – Educational English context

Education is compulsory for all children in England between the ages 5 to 16. Primary schools cover ages 5 to 11 (Reception to Year 6). At age 11, there is a transition to secondary schools, which cover ages 11 to 16 (Years 7 to 11) or 11 to 18 (Years 7 to 13). Pupils of PISA-taking age (15-year-olds) are typically within the same year group (Year 11). England’s National Curriculum emphasises traditional subject disciplines and knowledge. As argued by Greany and Earley (2021), the national curriculum rejects the move towards developing ‘21st century’ skills and competencies, which is apparent in many other school systems (e.g., Creese et al., 2016). The precise definition of ‘21st century’ skills is a contentious issue (Greany & Earley, 2021), however, the basic argument is that a broader set of cognitive and non-cognitive skills, such as critical thinking and collaboration, is nowadays needed for young people to thrive in a globally competitive marketplace (Ananiadou & Claro, 2009).

National curriculum tests (also known as SATs) in the subjects of English and mathematics are taken at the end of primary school, and General Certificate of Secondary Education (GCSE) exams in English, mathematics, and science (among a range of other subjects) are taken at the age of 16 (Greany & Earley, 2021). As has been argued by Baird et al. (2011, p. 14), “there is a great deal of angst about over-assessment in the English education system, with the phrase ‘assessment as learning’ having become associated with the fact that students are being taught to the test and are learning test materials”. In recent years, the national assessment curriculum has been described as having arguably more influence on schools than the National Curriculum, due to the ways in which assessment outcomes are used to hold schools accountable (Greany & Earley, 2021). Compared to the rest of the United Kingdom, England takes a somewhat different approach to external accountability demands (e.g., publishing annual school performance tables) and other more recent policy developments, i.e., academies and free schools programme, which has not been introduced elsewhere within the UK (Jerrim & Shure, 2016).

Furthermore, it has been argued that PISA represents the only UK-wide assessment taken by a sample of pupils on a regular basis (Jerrim, 2021). PISA tests students’ skills in reading, mathematics, and science, which are also subjects assessed in the national GCSE exams in England. One of the main aims of the PISA study, is to evaluate the extent to which 15-year-old students are prepared to meet the challenges of today’s knowledge societies and can apply the knowledge and skills they have learned and practised at school in unfamiliar settings (OECD, 2017a). Although there is a strong correlation between PISA scores and GCSE grades, there are also important differences between PISA tests and GCSEs (Jerrim & Shure, 2016). Some of these differences are presented in Table 2.1; for an overview see the National Report for PISA 2015 results in England (Jerrim & Shure, 2016).

Table 2.1. Main differences between PISA 2015 and GCSE exams

	PISA 2015	**GCSE exams**
Skills targeted	Ability to apply knowledge to meet real-life challenges	Knowledge of curriculum content areas
Test administration time	November/December 2015	May/June 2016
Test administration mode	Computer-based assessment	Paper-based assessment
Test questions	More reading demand for science and mathematics	Less reading demand for science and mathematics
Stakes of test	Low stakes, students receive no feedback about their performance	High stakes, students receive a grade that can impact future educational options and career

Notes: Source: National Report for PISA 2015 Results for England (Jerrim & Shure, 2016).

Taking what is also called a literacy perspective (Turner & Adams, 2007), PISA does not strive to assess curricula directly, but rather assesses students' ability to apply knowledge in new situations, avoiding a heavy emphasis on the assessment of factual recall and information retrieval (Baird et al., 2011). A consequence of this approach is having tests with a relatively high proportion of text-heavy questions with open-ended formats (Baird et al., 2011). Comparing the style of mathematics and science items used in PISA 2000 and 2003, GCSE, and Key stage 3 tests in England, Ruddock et al. (2006) found that the amount of reading required in PISA, differentiates it from the other assessments and is something that students in England would not be familiar with. In addition to that, a high level of numeracy is demanded in PISA reading tests, along with a demand to interpret diagrams (Ruddock et al., 2006), both of which are argued to be features of reading assessment that students in England are not familiar with (Baird et al., 2011).

2.4 Theoretical background for validity

Validity has been a fundamental concept in social science research, often referred to as the most critical consideration in developing and evaluating tests (Maul, 2018). In the following sections, a brief historical review on the concept of validity is offered. First, early perspectives on validity theory prior to the 1950s are presented. Next, the introduction of construct validity in the early 1950s is discussed, followed by a unified validity theory, due primarily to the work of Samuel Messick. Messick's (1989) unitary framework of validity is further described in more detail, since it guides the validation process followed in this thesis. Finally, current issues and controversies around the use of social consequences in the validation process are summarised.

2.4.1 Early perspectives of validity: Prior to the 1950s

Educational and psychological testing became prominent in the early 1900s with theorists sought to provide an account of validity (Hathcoat et al., 2018). In early perceptions of validity, the concept was understood in terms of the correlation between test scores and a criterion measure (Maul, 2018). Concerns about the connection between test content and validity, raised by educational researchers, influenced those early perceptions (Hathcoat et al., 2018). In short, validation required a criterion measure assumed to provide the 'real' value of the attribute of interest (Shaw & Crisp, 2011, p. 14).

2.4.2 Criterion, content, and construct validity: Early 1950s

The concept of validity was refined during the 1950s to include the ability of a test to predict future performance with respect to external criteria, content area, or a theoretical construct (Shaw & Crisp, 2011). Validity has been broken into three distinct types: criterion, content, and construct validity (Messick, 1987). **Criterion validity** is evaluated by comparing the test scores with one or more external variables (criteria) considered to correlate (Messick, 1987). It comprises two subtypes: concurrent and predictive criterion-related validity. The former indicates the extent to which the test scores estimate an individual's present standing on the criterion, while the latter indicates the extent to which an individual's future level on the criterion is predicted from prior test performance (Messick, 1987). **Content validity** is evaluated by showing how well the test content samples the subject about which conclusions are to be drawn (Messick, 1987). Evidence for content validity, however, has been argued to be subjective and confirmatory, making it difficult to justify conclusions about interpretation of test scores (Shaw & Crisp, 2011). **Construct validity** is evaluated by determining the degree to which certain explanatory constructs account for performance on the test (Messick, 1987). Essentially, this type of validity attempted to make a link between observed performance on an assessment and pre-conceived theoretical explanations (Shaw & Crisp, 2011).

2.4.3 Unified validity theory: 1980s-1990s

Samuel Messick offered a new perspective on validity as a unified concept, which reflected a significant shift from previous viewpoints (Maul, 2018). Specifically, as defined by Messick, validity is "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13). Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores for consequential action (Messick, 1995). Messick's main argument for abandoning the distinct types of validity view was that the previous conception was fragmented, and it did not incorporate the value of score meaning and social consequences of score use (Ghaderi, 2018).

In a unified validity view, score-based inferences depend on the social consequences of the testing, and therefore, these cannot be ignored in considerations of validity evidence (Messick, 1987). As a unified concept, validity does not diminish content or criterion validity evidence, but instead subsumes them to build a robust validity argument (Shaw & Crisp, 2011). Validation, therefore, is a continuous and open-ended process based on the accumulation of evidence from multiple sources (Ghaderi, 2018). In 1999, a unitary validity definition was endorsed by the American Educational Research Association, the

American Psychological Association, and the National Council on Measurement in Education (Ghaderi, 2018). Since then, it has been incorporated in the Standards for Educational and Psychological Testing, a document that provides guidance and addresses professional and technical issues of test development and use in education, psychology, and employment (American Educational Research Association et al., 2014).

2.4.3.1 Messick's unitary framework of validity

Validity evaluation rests on four bases (Table 2.2), which are: "(1) an inductive summary of convergent and discriminant research evidence that the test scores are interpretable in terms of a particular construct meaning, (2) an appraisal of the value implications of that interpretation, (3) a rationale and evidence for the relevance of the construct and the utility of the scores in particular applications, and (4) an appraisal of the potential social consequences of the proposed use and of the actual consequences when used" (Messick, 1980, p. 1023). Putting these four bases together, validity can be represented in terms of two interconnected facets linking the source of the justification (evidential or consequential) to the function or outcome of the testing (interpretation or use) (Messick, 1980). The ambiguity of these distinctions derives from the fact that the framework tries to cut through what indeed is a unitary concept (Messick, 1987).

Table 2.2. Messick's facets of validity (Messick, 1980)

	Test Interpretation	**Test Use**
Evidential Basis	Construct Validity	Construct Validity + Relevance/Utility
Consequential Basis	Value Implications	Social Consequences

Furthermore, Messick specified six distinguishable aspects of construct validity as a means of addressing central issues implicit in the notion of validity as a unified concept (Brussow, 2018). These are content, substantive, structural, generalisability, external and consequential aspects of construct validity (Messick, 1995). In effect, these six aspects function as general validity criteria or standards for all educational measurement, including performance assessments (Messick, 1995). These are briefly described next:

1) The **content** aspect includes evidence of content relevance, representativeness, and technical quality (Messick, 1995). Validity evidence can be obtained from an analysis of the relationship between the content of a test and the constructs it is intended to measure (American Educational Research Association et al., 2014).
2) The **substantive** aspect refers to theoretical rationales for the observed consistencies in test responses, along with empirical evidence that the theoretical processes are engaged by respondents in the assessment tasks (Messick, 1995). Analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of the performance or response engaged in by test takers (American Educational Research Association et al., 2014).
3) The **structural** aspect appraises the extent to which the internal structure of the assessment reflected in the scores is consistent with the structure of the construct domain (Messick, 1995). Analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based (American Educational Research Association et al., 2014).
4) The **generalisability** aspect examines the extent to which score properties and interpretations generalise appropriately to and across population groups, settings, and tasks (Messick, 1995). Evidence relating to the generalisability aspect may focus, for example, on the invariance of calibrations across measurement contexts (Wolfe & Smith, 2007b).

5) The **external** aspect includes convergent and discriminant evidence (Messick, 1995). Analyses of the relationship of test scores to variables external to the test, hypothesised to measure the same construct, related or different constructs, can provide a source of validity evidence (American Educational Research Association et al., 2014).
6) The **consequential** aspect appraises the implications of score interpretation as a basis for action as well as the actual and potential consequences of test use (Messick, 1995). While not all consequences can be anticipated, in some cases factors such as prior experiences in other settings offer a basis for anticipating unintended consequences (American Educational Research Association et al., 2014).

2.4.3.2 Social consequences: current issues and controversies

Social consequences, as a source of validity evidence, largely derived from Messick's unified validity framework and the proposition that social consequences inform, and are informed by, the meaning attributed to test scores (Hathcoat et al., 2018). Consequences can be positive or negative and intended or unintended, and their significance depends on the type (formative or summative) and the stakes (hight or low) that characterise an assessment (Ghaderi, 2018).

Although Messick's unified validity definition has been widely endorsed, the use of social consequences as a source of validity evidence have been a controversial topic. Opponents argue that social consequences are beyond the scope of a validation study and should be deferred to policy makers instead (Ghaderi, 2018). Another line of criticism suggests that since consequential evidence deals with ethical (instead of measurement) considerations, it should not be considered as part of validity evidence (Brussow, 2018). Finally, opponents of the use of social consequences as validity evidence, claim that, by considering social consequences as part of the validity concept, leads to further confusion surrounding validity (Brussow, 2018). Two issues of the journal *Educational Measurement: Issues and Practice* in 1997 and 1998 targeted the topic of social consequences as validity evidence, illustrating the importance of this debate (Shaw & Crisp, 2011).

Advocates of considering social consequences as part of validity evidence argue that consequences reflect the soundness of test-based decisions (Ghaderi, 2018). As far as adverse consequences are concerned, as Messick (1989, 1995) explained, the primary concern is that any negative impact on individuals or groups should not derive from any source of test invalidity, such as construct underrepresentation or construct-irrelevant variance. Furthermore, consideration of the social consequences is an important ethical consideration when designing a validation study (Brussow, 2018). Although, the use of social consequences remains a controversial topic among validity theorists (Hathcoat et al., 2018), the current Standards for Educational and Psychological Testing suggest that intended consequences should be investigated as a validity issue (American Educational Research Association et al., 2014).

2.4.4 Contemporary views toward validation

Messick's unified validity theory has remained influential since its introduction and is still argued to be the dominant validity theory in the literature on educational assessment and measurement (Maul, 2018). Building on Messick's (1989) unitary view of validity, scholars such as Kane (2006) proposed an argument-based approach to validation, which involves the evaluation of an argument aimed at defending the appropriateness of a test for a particular use (Maul, 2018).

2.5 Current validity issues and criticisms concerning the PISA study

Due to PISA's far-reaching political influence, the investigation of the measurement tools used, and the validity of the derived measures, are argued to be of particular importance (Hopfenbeck, 2016). Various validity aspects surrounding the PISA study have been questioned both within the UK and internationally (Hopfenbeck et al., 2018; Jerrim, 2021). Lines of criticism raised in the literature include validity issues with the PISA questionnaire instruments, materials translation, the measurement model used for producing PISA scores, student sampling, and consequential validity (Baird et al., 2011; Goldstein, 2004).

A major criticism relates to the way that results have been interpreted and used to introduce educational reforms (Ercikan et al., 2015; Shiel & Eivers, 2009). In a recent literature review on PISA (Hopfenbeck et al., 2018), a substantial number of articles were found to advise policy makers and researchers to be cautious about using PISA data as a means for valid comparison or for informed policymaking. In addition, several authors have criticised PISA questionnaires (e.g., Caro et al., 2014; El Masri et al., 2016; Hopfenbeck & Maul, 2011). For example, Caro et al. (2014) raised concerns about the differential meaning of the cultural, social, and economic constructs across the countries making cross-cultural comparisons difficult to establish. In another study, Hopfenbeck and Maul (2011) found that the scales of self-report questionnaires led to invalid responses largely due to poor questionnaire design and language ambiguity. Additionally, questions have been raised about cross-country differences in translation and interpretation of PISA's test material (El Masri et al., 2016). Regarding the sampling approach followed by PISA, Anders et al. (2021) showed how a combination of low response rates and high exclusions lead to serious questions surrounding the representivity of the PISA 2015 data for Canada.

Another line of criticism concerns the scaling methodology adopted for the production of PISA student test scores (Goldstein, 2017; Jerrim, Parker, et al., 2018). Various authors described the process as opaque, which may have implications for subsequent use of the data (Goldstein, 2017; Jerrim, Parker, et al., 2018). Others have suggested that the Rasch model (Rasch, 1960) used in the PISA study until 2015 is overly simplistic and does not fit the data well (Kreiner & Christensen, 2014). Following such criticism, a more complex model, the two-parameter logistic model, was used to construct the scale scores in PISA 2015 (OECD, 2017c). However, a recent study analysing PISA 2015 student data for reading, mathematics and science highlighted that some of the criticisms made of PISA's past scaling methodology are unjustified (Jerrim, Parker, et al., 2018). Specifically, it found that cross-country comparisons of educational achievement did not really change when a more complex methodology, instead of the Rasch model, was used for scaling (Jerrim, Parker, et al., 2018).

Issues surrounding transparency of reporting have been recently raised about the PISA results. Specifically, in England, a non-response bias analysis was produced, but not published by the OECD (Jerrim, 2021). Using PISA 2018 data for all four nations of the UK, Jerrim (2021) found evidence of an upward bias in the data for England and Wales, with lower achievers systematically excluded from the sample. Following that, the UK Statistics Authority was called upon by the author to conduct a review of the PISA 2018 data for the UK, with the issue getting attention from the national media (e.g., Coughlan, 2019). Concerns have also been raised regarding the switch between paper and computer assessment (Jerrim, 2016; Jerrim, Micklewright, et al., 2018).

Regarding the PISA 2015 CPS assessment, several issues have been raised in the literature so far, which are covered in detailed in the following results chapters. As mentioned in Chapter 1, computer-simulated partners in the PISA 2015 CPS assessment were programmed to represent team members with different roles, attitudes, levels of competence, as well as behaviour to vary the CPS situation the students are confronted with (OECD, 2017a). This approach was preferred due to the high degree of control and standardisation required by the assessment. However, whether computer-simulated partners can be designed to reliably mimic realistic conversational partners, or the extent to which interacting with computer-simulated partners generalises to interacting with human partners, remains an open question (Webb & Gibson, 2015).

The approach of replacing real humans with computer-simulated partners in the groups has been criticised for lacking authenticity and deviating from ecologically valid CPS activities (Graesser et al., 2018; Rosen, 2015; Scoular et al., 2017; Siddiq & Scherer, 2017). Thus, in many cases, the validity of the CPS competence measures derived from such constrained assessments has been questioned (Rosen & Foltz, 2014; Scoular & Care, 2020). From a conceptual perspective, a limited range of information from CPS tasks has been released for public view, e.g., items for only one of the CPS tasks was released (De Boeck & Scalise, 2019). Finally, researchers using the PISA 2015 CPS data do not have the information to examine what is happening during the assessment, since descriptions of the possible actions are not available in the data set (De Boeck & Scalise, 2019).

2.6 Closing remarks on Chapter 2

This chapter aimed to clarify the ‘modern’ conceptualisation of validity of assessments such as PISA, and particularly to situate these in the context of their policy/political consequences. This will provide the basis for evaluating the criticisms of PISA’s validity in the previous literatures in Chapters 4 and 5 (especially of CPS) and motivate the empirical studies (and their adopted method/ologies) in Chapters 6 and 7. Finally, it will support the synthesis of the whole approach of the thesis to validation in the international and English context. In the following chapter (Chapter 3), the methodological rationale and methods of the thesis are presented.

Chapter 3 Methodological Rationale and Methods

3.1 Summary of the chapter

Guided by the research aims and research questions, this thesis includes two systematic literature reviews (Chapters 4 and 5/Research papers 1 and 2), and two consecutive empirical phases in which different methods are applied on different data (Chapters 6 and 7/Research papers 3 and 4). This chapter discusses the methodology and research design relevant to this thesis and presents in a detailed way information related to the methodology section of each research paper (presented in Chapters 4-7). It explains why a sequential mixed methods design is chosen and the philosophical underpinning adopted while considering an underlying belief in the complementarity of different research approaches. For the two systematic literature reviews (Chapters 4 and 5), the literature review methodologies adopted are discussed in detail. Specifically, this chapter presents the following steps that guided the reviews: article search approach, inclusion criteria, selection of articles, and analysis. For the first empirical phase (Chapter 6), a description of the Programme for International Student Assessment (PISA) 2015 dataset is provided, including information about sampling, test design, features of collaborative problem solving (CPS) assessment and a sample item. Then, the analytical approach is discussed, including a brief description of the family of item response theory (IRT) models that have been typically used for the analyses of PISA data, the advantages of Rasch measurement, and the description of validation process using the Rasch model. For the second empirical phase (Chapter 7), procedures of qualitative data collection and data analysis are discussed, including participant selection, cognitive interview protocol design, and grounded theory analysis. The chapter concludes with strategies to address issues with reliability and trustworthiness.

3.2 Research aims and research design

The main aims of this thesis were to understand what CPS competence means, its conceptualisation and operationalisation, and how these inform the validation of the PISA 2015 CPS assessment and derived CPS competence measures. Although the topic of the validity of PISA measures has attracted lots of attention, research on the topic of PISA 2015 CPS competence measure’s validity is relatively weak.

In addressing the research aims, this thesis is guided by the following research questions and sub-questions:

RQ1. How has “collaborative problem solving” been conceptualised and operationalised in the educational research community?

RQ1.a: How are the variations in the CPS conceptualisations explained by diverse research purposes?

RQ2. How has students’ CPS competence been assessed using computer-simulated, scenario-based assessment tasks in educational research studies?

RQ2.a: What are the existing assessments of students’ CPS competence and their characteristics (e.g., subject domain, task design features)?

RQ2.b: Which facets of CPS competence do the assessments measure?

RQ2.c: What strategies for validating CPS competence measures are reported?

RQ3. What are the strengths and limitations of measurement validity of the CPS competence measure for England based on PISA 2015 data?

RQ3.a: To what extent is a hypothetical three-dimensional structure of ‘establishing and maintaining shared understanding’, ‘taking appropriate action to solve the problem’, and ‘establishing and maintaining team organisation’ measures supported empirically by the PISA 2015 data for CPS assessment in England?

RQ3.b: To what extent are the constructed measures of CPS competence invariant across gender?

RQ3.c: How are the constructed measures of CPS competence related to other relevant (collaboration and performance) constructs?

RQ4. What does the PISA 2015 CPS assessment actually measure according to student perspectives?

RQ4.a: How do students comprehend the CPS assessment items and how do they explain their answers to them?

RQ4.b: What are the implications for the external validity of the CPS assessment?

RQ1 and RQ2 are of a more conceptual nature and are answered in two systematic literature reviews presented in Chapters 4 and 5 respectively. RQ3 is answered in the first empirical phase (Chapter 6) through conceptualisation/operationalisation and validation (building on the two systematic literature reviews and PISA methodology) as well as further statistical modelling of student CPS competence measures. I use the PISA 2015 student dataset for England to investigate the hypothesised multidimensional structure of CPS competence construct, as well as its relation to other relevant (collaboration and performance) constructs and background variables. RQ4 is answered in the second empirical phase (Chapter 7) using the qualitative method of cognitive interviewing to collect data from secondary school students in England. In this sequential phase, students' response processes are explored to better understand what the CPS competence construct derived from the PISA study means by interpreting the students' own discourses as their expressions of understanding and competence.

3.3 Mixed methods research

3.3.1 Pragmatism as a philosophical position

The philosophical underpinning of pragmatism considers the ways in which a pragmatic stance is adopted and linked with the current mixed methods research project. Pragmatism is argued to put aside ontological and epistemological debates about what and how we can know the social world (Tashakkori & Teddlie, 1998). The theoretical perspective of pragmatism derived from the work of Charles Sanders Peirce, William James, Georg H. Mead, and John Dewey, who were interested in examining practical consequences and empirical findings to help in deciding how to better understand real-world phenomena (Johnson &

Onwuegbuzie, 2004). It helps to focus research on the research question and use philosophical and methodological approaches that work best for the problem under study (Teddlie & Tashakkori, 2009). Following a pragmatist stance, this study engages with multiple and diverse research questions and acknowledges that the validation of a CPS measure can be carried out in relation to different types of validity evidence including evidence derived from student interviews.

3.3.2 Definition of mixed methods research

Several definitions of mixed methods research are available in the literature (e.g., Johnson et al., 2007; Teddlie & Tashakkori, 2009). Tashakkori and Creswell have defined mixed methods as 'research in which the investigator collects and analyses data, integrates the findings, and draws inferences using both qualitative and quantitative approaches or methods in a single study or a program of inquiry' (2007, p. 4). Mixed methods research has emerged as an alternative to the traditional dichotomy between qualitative and quantitative research traditions in the social and behavioural sciences (Sammons & Davis, 2017; Teddlie & Sammons, 2010). Researchers have rejected the arbitrary opposition of qualitative and quantitative approaches as competing alternatives that they have to make a forced choice between (Sammons & Davis, 2017).

The need to move research design beyond the traditional and often oppositional quantitative or qualitative dichotomy has been highlighted in the literature (Creswell, 2003). Towards that direction, the field of mixed methods research has been argued to move past the paradigm wars by offering a logical and practical alternative on how both paradigms can be used together in a study (Johnson & Onwuegbuzie, 2004). Therefore, mixed methods research bridges the gap between the quantitative and qualitative oppositions and is used to maximise the strengths and minimise the weaknesses of the two (Johnson & Onwuegbuzie, 2004). Following the publication of the Handbooks of Mixed Methods Research (Tashakkori & Teddlie, 2003, 2010), mixed methods has been recognised as a third and alternative methodological approach (Sammons & Davis, 2017). Its increasing popularity is attributed to its flexibility in addressing multiple and diverse research questions through qualitative and quantitative techniques (Teddlie & Sammons, 2010).

3.3.3 Methodological rationale for mixed methods research

Although some authors have criticised mixed method approaches due to issues of compatibility in combining qualitative and quantitative methods (Denzin & Lincoln, 1994; Howe, 1985; Morgan, 2007), a growing number of researchers have argued that mixed methods research provides insights and understanding that might be missed when using only a single method (Creswell & Plano Clark, 2007; Johnson & Onwuegbuzie, 2004; Tashakkori & Teddlie, 2010). For example, some of the strengths of adopting mixed methods research, as presented by Johnson and Onwuegbuzie (2004), include answering a broader and more complete range of research questions since the researcher is not confined to a single method or approach. The appeal of mixed methods research lies in its ability to combine both numeric findings and stories to generate new knowledge (Teddlie & Sammons, 2010). Mixed methods research is an inclusive, pluralistic, creative, and complementary form of research, which rejects restricting or constraining researchers' choices in answering research questions (Johnson & Onwuegbuzie, 2004). As argued by Johnson and Onwuegbuzie (2004), researchers should collect multiple data using different methods based on complementary strengths and nonoverlapping weaknesses.

The purpose of adopting a mixed methods design in this study was to seek benefits of complementarity to make stronger inferences. Some of the strengths of quantitative research, that are relevant for this study, include: testing and validating already constructed theories about how phenomena occur, obtaining data that allow quantitative predictions to be made, relatively less time-consuming data analysis, and studying large numbers of people (Johnson & Onwuegbuzie, 2004). Similarly, strengths of qualitative research, relevant for this study include: obtaining data that are based on the participants' own categories of meaning, studying a limited number of cases in depth, describing complex phenomena, and determining how participants interpret "constructs" (Johnson & Onwuegbuzie, 2004, p. 20).

3.3.4 Sequential mixed methods design

Researchers have created a range of typologies to describe and classify mixed methods research designs (Creswell, 2003; Creswell & Plano Clark, 2007; Leech & Onwuegbuzie, 2009; Teddlie & Tashakkori, 2006). To construct a mixed methods design for this thesis, I follow Johnson and Onwuegnuzie's (2004) definition of mixed methods research as: "the class of research where the researcher mixes or combines quantitative and qualitative research techniques, methods, approaches, concepts or language into a single study" (p. 17). In accordance with the relationship between each research papers' purpose, this thesis can be described as a mixed methods sequential explanatory design (Creswell & Plano Clark, 2007). As illustrated in Figure 3.1, a four-phase mixed methods design is followed, with each phase being a Chapter/research paper of the thesis. The first two phases are two systematic literature reviews, which are primarily qualitative, being conceptual before they really get into developing quantitative summaries. The latter two phases are empirical with qualitative data helping explain and build upon quantitative results, both informed by the systematic reviews.

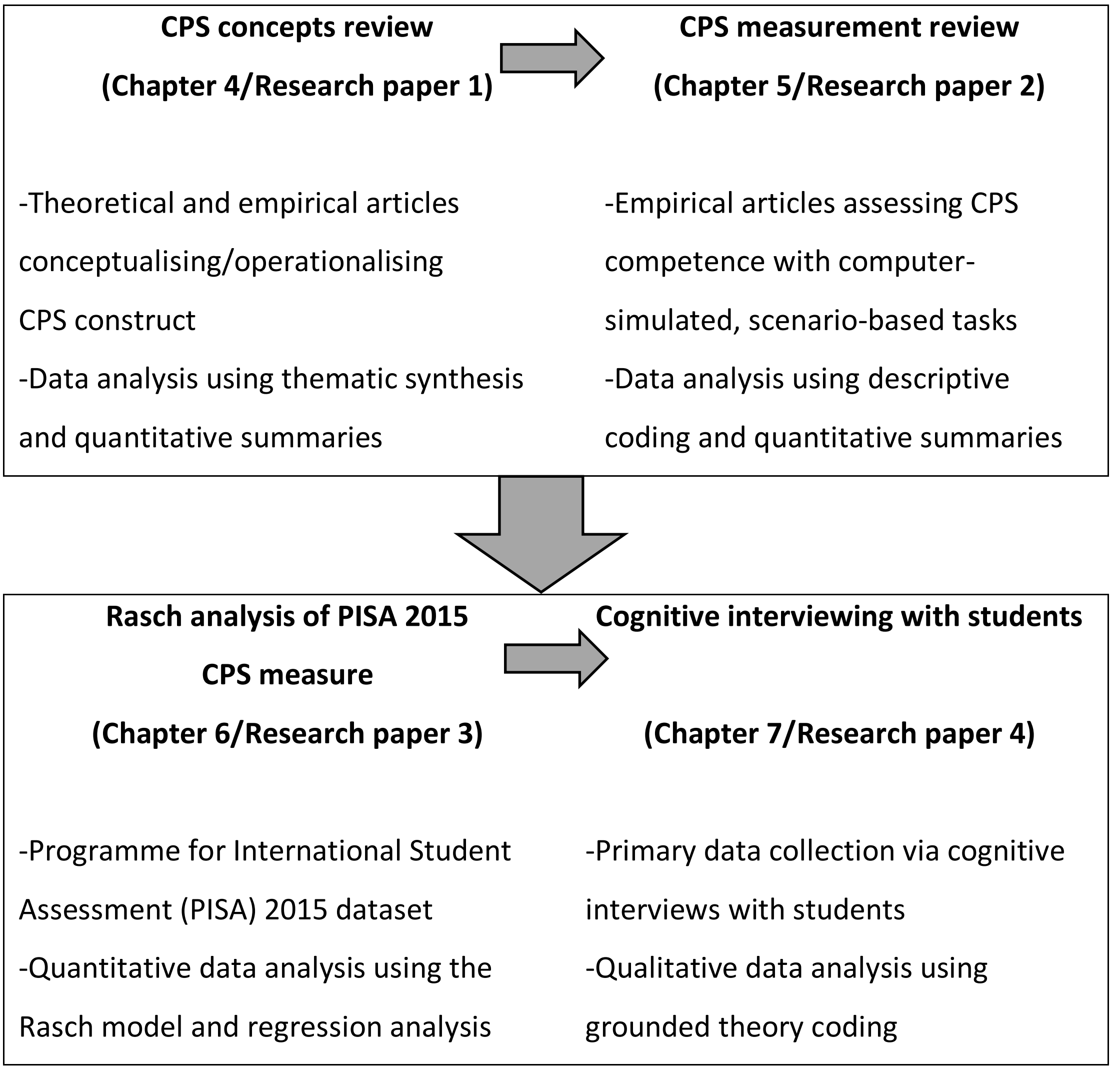

CPS concepts review
(Chapter 4/Research paper 1)
-Theoretical and empirical articles conceptualising/operationalising CPS construct
-Data analysis using thematic synthesis and quantitative summaries
CPS measurement review
(Chapter 5/Research paper 2)
-Empirical articles assessing CPS competence with computer-simulated, scenario-based tasks
-Data analysis using descriptive coding and quantitative summaries
Rasch analysis of PISA 2015 CPS measure
(Chapter 6/Research paper 3)
-Programme for International Student Assessment (PISA) 2015 dataset
-Quantitative data analysis using the Rasch model and regression analysis
Cognitive interviewing with students
(Chapter 7/Research paper 4)
-Primary data collection via cognitive interviews with students
-Qualitative data analysis using grounded theory coding

Figure 3.1. Diagram of sequential mixed method design

Notes: Arrows stand for sequential (following Johnson & Onwuegbuzie, 2004).

Findings related to the conceptualisations of CPS from Chapter 4 (CPS concepts review) are used as a backdrop for Chapter 5 (CPS measurement review) to examine the assessment of CPS competence with the use of computer-simulated, scenario-based assessment tasks. Additionally, the dangers in privileging only research evidence using one category of conceptualisations (i.e., CPS competence) is discussed in Chapter 4, with a view to taking this work further in Chapter 5. Finally, Chapter 5 discusses the limitations in current assessments of CPS competence, and the gaps in the validity evidence concerning CPS competence measures, with a view to taking this work further in Chapters 6 and 7.

As a result of the criticisms around the validity of the PISA 2015 CPS assessment, presented in detail in Chapters 4 and 5 (systematic reviews), Chapter 6 (Rasch analysis of PISA CPS measure) examines the dimensionality and subsequently aspects of validity of the PISA 2015 CPS competence measures. Chapter 6 uses the available secondary data (PISA 2015), which allowed me to start exploring validity before collecting qualitative data and before moving into in-depth data analysis. The measure validation as well as the descriptive analysis allowed the questioning of validity aspects of the CPS competence measures. Although this phase did not explore possible explanations of students' response processes, it provided evidence (e.g., item fit statistics and item difficulties) that could be further understood and explained by the subsequent collection and analysis of qualitative data.

In Chapter 7 (Cognitive interviewing with students), the validity of the PISA 2015 CPS competence assessment is explored using more in-depth qualitative methods, building on the quantitative results derived from Chapter 6. Chapter 7 is also conceptually informed by the results of Chapter 4 and 5 (systematic reviews). In the empirical phases of this study, quantitative data were analysed and interpreted first, and then qualitative data were subsequently collected and analysed. Equal weight was given to methods since they play an equally important role in addressing the research aims. The two datasets were analysed separately in the results section of the research papers (Chapter 6 and 7) and were then synthesised in the discussion (Chapter 8).

3.4 Methodology for CPS concepts review (Chapter 4)

3.4.1 Overview

The CPS concepts review (Chapter 4) is concerned with systematically examining how CPS has been conceptualised in recent empirical and theoretical educational research. To get a broader understanding of the literature, I examine how researchers defined the construct, what theories they used to support their definitions and how they operationalised it in the empirical part of their work (where appropriate).

Gough, Oliver, and Thomas (2016, p. 2) define a systematic review as “a review of existing research using explicit, accountable rigorous research methods”. For them, a review of research is a form of research in itself, and a systematic review is ‘systematic’ in the same way that any empirical research needs to be systematic and transparent, so that the results can be interpreted and assessed in the light of how they were produced. A pre-defined procedure proposed by Gough et al. (2016) and Petticrew and Roberts (2006) to guide systematic reviews including the following steps was employed: developing research questions, determining the types of publications to be located, carrying out a comprehensive literature search, formulating inclusion criteria, appraising study quality, extracting data and synthesising.

3.4.2 Literature database: article search approach and inclusion criteria

To gather relevant evidence, an electronic search was conducted in four scientific databases: SCOPUS, Web of Science, Education Resources Information Centre (ERIC), and British Education Index (BEI). These databases were chosen as they offer an extensive coverage of research literature in the social sciences and two of them (i.e., ERIC, BEI) are relevant to educational research. This review was focussed on the field of education, since students emerging from schools into the workforce and public life are expected to be able to work in teams to solve diverse problems (Rosen & Foltz, 2014). The search was first conducted in January 2019 and was updated in April 2020[2] to include the most recent publications. The terms “collaborative problem solving” and “student(s)”, “pupil(s)” or “learner(s)” were used in the search.

[2] The search was also updated on the 21st of July 2022 in the Scopus database using the same keywords and restrictions to check for the number of new articles published between 2021 and 2022 (see Appendix 2). The analysis of these was beyond the scope of this study, since the aim of the conceptual review was to inform the empirical phases following the research design. Therefore, integration with the results was not considered at this time, but future research work may consider extending the search.

The search targeted research literature which focused specifically on the CPS construct rather than related terms, such as “teamwork”, and “problem solving”. Boolean operators were employed to combine the key terms as follows: “collaborative problem solving” AND (student* OR pupil* OR learner*), making the search specific to student populations. The search terms were applied to the fields of title, abstract, and keywords.

Articles were included based on the following criteria:

- peer-reviewed journal articles,
- published between 2000 and 2020,
- written in the English language,
- full text available,
- referred to student populations from various educational settings (from reception to higher education),
- were concerned about CPS in the field of education,
- provided a clear definition/conceptualisation/framework of CPS.

The publication period was restricted to the last two decades aiming to map current literature of the 21st century. Search was limited to articles published in peer-reviewed journals to assess study quality, although it is recognised that this has its own limitations as a quality check criterion (Alexander, 2020). A publication was retained for review when CPS-related research formed part of the content and focus of the article, meaning that CPS was not used merely as an example among other aspects of learning, and CPS was not used as a term specific to a field other than education. In this way, only articles with a focus on CPS, targeting student populations and providing explicit evidence for their conceptualisation were considered as relevant for inclusion in the review.

3.4.3 Selection of articles

To screen and select articles to be included in the literature review, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) method (Moher et al., 2009) was used (Figure 3.2). Publication records were managed in the systematic review software EPPI-Reviewer 4 (Thomas et al., 2010). The article search yielded 702 articles in total. First, duplicate entries were removed across databases resulting in 374 unique articles to be screened. Screening was conducted in two rounds: 1) applying the inclusion criteria to titles and abstracts, and 2) applying the inclusion criteria to the full text of the remaining articles. If a decision for inclusion could not be made by reading only the title and abstract, then the article was included in the sample for further screening on the full text. Following the first screening round, a total of 218 articles were deemed relevant to be screened by full text. Following the second screening round, 54 articles were selected to be analysed and reviewed (see Figure 3.2; the full reference list is included in Chapter 4).

To ensure that important and relevant research evidence has not been missed, a targeted search was conducted. Specifically, practices of referential backtracking and researcher checking were used to look beyond the results of the database searches (Alexander, 2020). This involved examining the reference lists of existing literature reviews (i.e., Cukurova, Luckin, & Baines, 2018; Graesser et al., 2018; Oliveri et al., 2017). Finally, the publication records of authors, who were found to frequently contribute to the topic, were searched to determine if additional articles warranted review. This resulted in identification of five additional articles, bringing the total to 59 articles for review. Randomly chosen articles (n = 40) were reviewed by the second author and there was a 95% match in the ratings (i.e., inclusion and exclusion) of articles by both reviewers. Any disagreements were resolved through discussion.

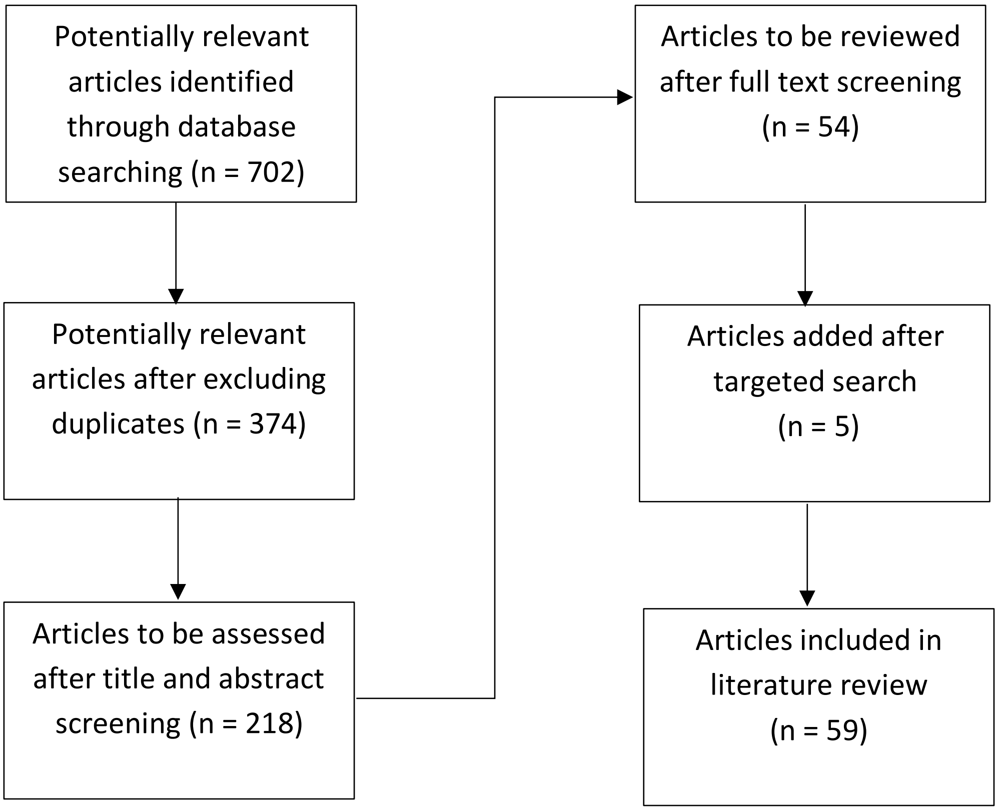

Potentially relevant articles identified through database searching (n = 702)
Potentially relevant articles after excluding duplicates (n = 374)
Articles to be assessed after title and abstract screening (n = 218)
Articles to be reviewed after full text screening (n = 54)
Articles added after targeted search (n = 5)
Articles included in literature review (n = 59)

Figure 3.2. Flow diagram of systematic literature review selection

3.4.4 Analysis

To make sense of the literature, information is summarised in a form that can be easily viewed, analysed, and managed, in other words, the literature is coded. Applying descriptive codes to articles enables the description or mapping of the size and nature of the literature before examining it in depth (Gough et al., 2016). Throughout the analysis, the unit of analysis was the article, and coding was based on the following descriptive variables:

- General information: authors, year of publication, journal name, country affiliation of the institution of the first author, type of research (empirical or theoretical).

- Research design (for empirical research only): evidence type, research aims, educational setting, sample size, curriculum area, group composition.

Thematic synthesis was employed (Thomas & Harden, 2008) to analyse definitions of CPS (and the associated operationalisations) from each article. Following this method, analysis was performed in three stages, as shown in Table 3.1. Specifically, analysis started from free line-by-line coding, staying close to the data itself, and then moved towards descriptive and analytical coding, building up to more abstract conceptualisations (Thomas & Harden, 2008).

As a first step, each article was read to gain insight into how CPS was understood and used by the author(s), looking for clear statements about CPS definitions. These were extracted in the form of literal quotes, along with theories consulted or used, research purposes and methodology. The reason for extracting information about methodology is that I consider the way a construct was measured to reveal a lot about the concept as practised, adding, in this way, significantly to what the concept was defined to be. The extracted raw data was coded focusing on identifying defining features of CPS conceptualisation that appeared to be important for the authors. This first step remained closely attached to the data, using constant comparison between data that was given the same code to check consistency of interpretation (Thomas & Harden, 2008).

In step two, descriptive themes were constantly refined to reflect the various defining features of CPS conceptualisations. In the final step, I have gone beyond the content of the original articles by using the descriptive themes to answer my research questions (Thomas & Harden, 2008). The main result of this final step was the categorisation of articles into three different groups based on the distinct foci guiding their conceptualisation. A description of their purpose and operationalisation was also produced considering the information extracted from each article’s stated research purpose and methodology.

Table 3.1. Worked example of analysis procedure (italics refer to the first step of analysis, where main concepts were highlighted for coding)

Step 1: Quote of definition	Step 2: Analysis of defining features	Step 3: Purpose—operationalisation
Article A (Care, Scoular, et al., 2016): The focus in this work is on *the skill of the individual in the collaborative partnership*, as opposed to a focus on the collaborating pair. (…) The CPS framework presents a conceptual hypothesis about what skills might enable individuals *to solve problems that are so complex and require so many different resources that one individual alone cannot reach a solution*. In this, the required skills are *the capacity to enact a process*, where that capacity can range from relatively primitive to sophisticated. (…) The approach taken in this work is aligned with the hypothesis that *skills such as collaborative problem solving can be assessed, and taught*, in the mainstream educational system. This means that these *skills must be amenable to deconstruction into their contributing subskills or elements* to facilitate design of assessment tasks; and to identification of lower to higher competency in order to facilitate their teaching.	-Complex problem that requires interdependency, -Individual capacity to enact a process, -Latent construct, breaking it down into low-level observable behaviours	Purpose: development of educational assessments, measurement of student learning outcomes

Article B (Gu & Cai, 2019): Collaborative problem solving (CPS) resonates with *sociocultural theory* (Vygotsky, 1978) and so may appear as a *promising educational approach for equipping learners with understandings and skills* (…). In contrast to studies of individual problem solving, CPS research focuses on optimising the *benefits of social interaction for facilitating the cognitive development of participants*. (…) Therefore, the questions in this study arise from the needs to investigate the effects of semantic diagram tools on transaction costs during CPS processes and associated levels of deep understanding. (…) Each utterance of group chat messaging was coded as the unit of analysis. (…) In this study, deep understanding was evaluated in the pretest and posttest.	-Educational approach that aims to develop students' cognitive development -Focus on social interaction	Purpose: intervention evaluation, assessment of student learning outcomes

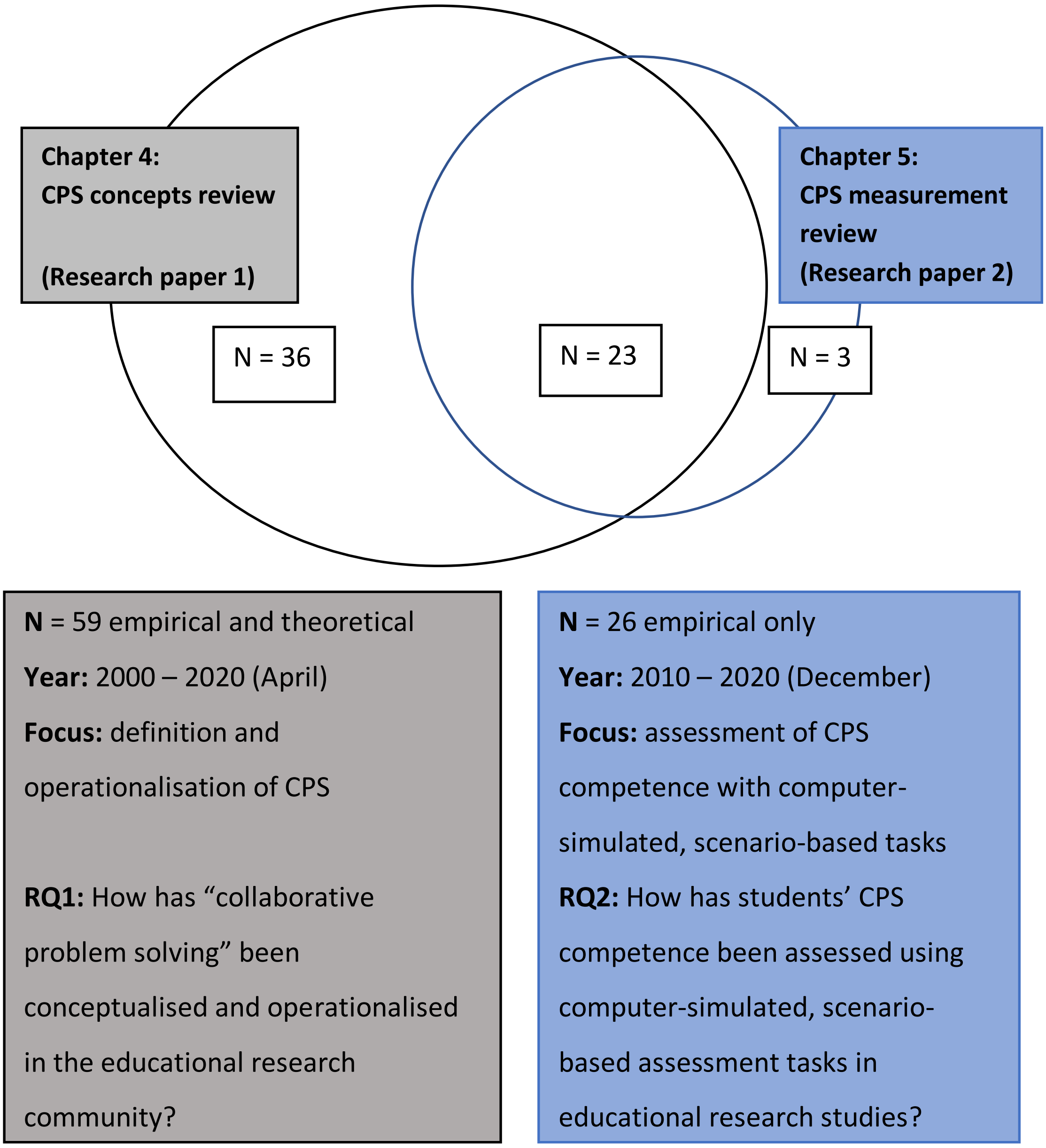
3.5 Methodology for CPS measurement review (Chapter 5)

3.5.1 Overview

The CPS measurement review (Chapter 5) takes a closer look at the methods of data collection (i.e., assessment instruments) and analysis of articles assessing students' CPS competence with the use of computer-simulated, scenario-based tasks. This is done for the purpose of developing a better understanding and a critique of existing assessments of CPS competence following a systematic literature review methodology (Gough et al., 2016; Petticrew & Roberts, 2006).

3.5.2 Connection to conceptual review

The connection between the CPS concepts review (Chapter 4) and CPS measurement review (Chapter 5) is depicted in Figure 3.3. The article search approach was the same in both reviews (i.e., databases and search terms), but the inclusion criteria differed to accommodate the distinct purposes of the two reviews as is evident from the research questions answered by each review (Figure 3.3).



N = 59 empirical and theoretical

Year: 2000 – 2020 (April)

Focus: definition and operationalisation of CPS

RQ1: How has "collaborative problem solving" been conceptualised and operationalised in the educational research community?

N = 26 empirical only

Year: 2010 – 2020 (December)

Focus: assessment of CPS competence with computer-simulated, scenario-based tasks

RQ2: How has students' CPS competence been assessed using computer-simulated, scenario-based assessment tasks in educational research studies?

Figure 3.3. Venn diagram of articles reviewed in the CPS concepts review and the CPS measurement review

While CPS concepts review included both empirical and theoretical articles (including reviews) to examine the various definitions and operationalisations of CPS, the CPS measurement review targeted only empirical articles that adopted a specific CPS conceptualisation and operationalisation. More specifically, for articles to be included in the second review they needed to define CPS as a student competence and assess it through computer-simulated, scenario-based tasks.

3.5.3 Literature database: article search approach and inclusion criteria

An electronic search was conducted in four scientific databases covering research in social sciences (including educational research): SCOPUS, Web of Science, Education Resources Information Centre (ERIC), and British Education Index (BEI). Assessments of CPS competence were previously reported in the fields of education, business, health and medicine (Oliveri et al., 2017). This review was focussed on the field of education, since students emerging from schools into the workforce and public life are expected to be able to work in teams to solve diverse problems (Rosen & Foltz, 2014). The search was first conducted in January 2019 and was updated in April 2020 and December 2020[3] to include the most recent publications.

Boolean operators were employed to combine the key terms as follows: "collaborative problem solving" AND (student* OR pupil* OR learner*), making the search specific to student populations. The search terms were applied to the fields of title, abstract, and keywords.

[3] The search was also updated on the 21st of July 2022 in the Scopus database using the same keywords and restrictions to check for the number of new articles published between 2021 and 2022 (see Appendix 2). The analysis of these was beyond the scope of this study, since the aim of the conceptual review was to inform the empirical phases following the research design. Therefore, integration with the results was not considered at this time, but future research work may consider extending the search.

Articles were included based on the following criteria:

- peer-reviewed journal articles,
- published between 2010 and 2020,
- written in the English language,
- full text available,
- reported data collection and analysis (primary or secondary),
- referred to student populations from educational settings (from reception to higher education),
- were concerned about the assessment of students' CPS competence in the field of education,
- provided a clear definition/conceptualisation/framework of CPS competence,
- developed (or used, for secondary data) at least one computer-simulated, scenario-based task as their assessment instrument.

To include timely and up-to date articles, the period between 2010 and 2020 was chosen to map current literature published in the last decade. Peer review was used as a quality check criterion, although it is recognised that this has its own limitations (Alexander, 2020). Additionally, a publication was retained for review when assessment of students' CPS competence formed part of the content and focus of the article, meaning that CPS was not used merely as an example among other learning outcomes or as a term specific to a field other than education. The selection of articles was restricted to those using computer-simulated, scenario-based tasks, excluding articles using solely other task types such as self-assessments or teacher evaluations, for two reasons: (i) a recent review (Oliveri et al., 2017) has already included examples of the above task types with the exception of computer-simulated, scenario-based tasks, and (ii) there is an increasing number of studies using computer-simulated, scenario-based tasks to draw inferences about students' CPS competence that have not been reviewed in a systematic manner to date.

3.5.4 Selection of articles

To screen and select articles to be included in the literature review, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) method (Moher et al., 2009) was used (see Figure 3.4). Publication records were managed in the EPPI-Reviewer 4 software (Thomas et al., 2010). The search yielded 665 articles in total. Duplicate removal resulted in 371 unique articles to be screened. Screening was conducted in two rounds: 1) applying the inclusion criteria to titles and abstracts, and 2) applying the inclusion criteria to the full text of the remaining articles. If a decision for inclusion could not be made by reading only the title and abstract, then the article was screened on full text. Following the first screening round, 93 articles were deemed relevant for full text screening. Following the second screening round, 22 articles were selected to be analysed and reviewed.

A targeted search involved examining the reference lists of articles included in the review to identify research that may fit within article search (Alexander, 2020). The reference lists of existing literature reviews (i.e., Graesser et al., 2018; Oliveri et al., 2017), were also used for referential backtracking even though those documents fell outside the inclusion criteria. Finally, the publication records of authors, who were found to frequently contribute to the topic, were searched. This resulted in identification of four additional articles, bringing the total to 26 articles for inclusion in this review (Figure 3.4). Randomly chosen articles (n = 40) were reviewed by the second author, resulting in a 95% match in the ratings (i.e., inclusion and exclusion) of articles by both reviewers. Any disagreements were resolved through discussion.

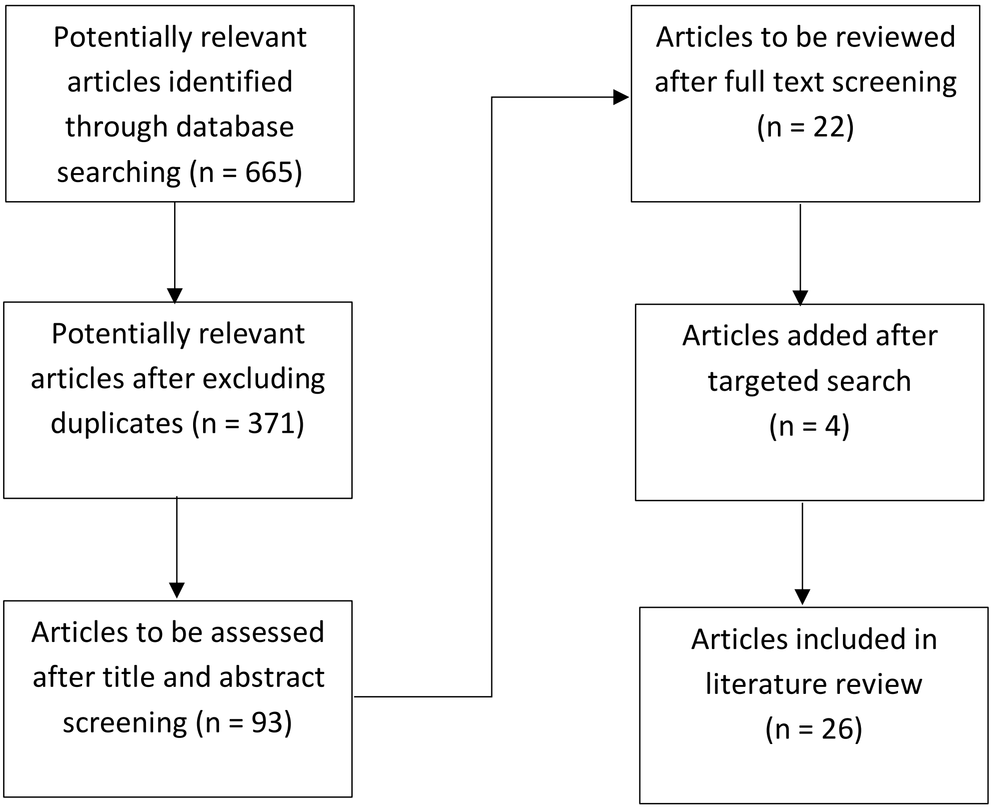

Potentially relevant articles identified through database searching (n = 665)
Potentially relevant articles after excluding duplicates (n = 371)
Articles to be assessed after title and abstract screening (n = 93)
Articles to be reviewed after full text screening (n = 22)
Articles added after targeted search (n = 4)
Articles included in literature review (n = 26)

Figure 3.4. Flow diagram of systematic literature review selection

3.5.5 Coding and data extraction tools

3.5.5.1 Assessment characteristics

To answer RQ2.a (What are the existing assessments of students' CPS competence and their characteristics (e.g., subject domain, task design features)?), the units of analysis were the assessments, and the information in the articles represented the source of data extracted. In the 26 articles included in this systematic review, 15 distinct assessments of CPS competence were represented. For some assessments, there was only one publication while others were reported in several publications. The following format for extracting relevant information about the assessments was developed:

- General information and research design: authors, year of publication, country(ies) of data collection, instruments of data collection, sample size, educational level.

- Task design features: group size, partner mode (i.e., human, or computer-simulated), subject domain, communication mode, scoring approach.

3.5.5.2 Facets of CPS competence

To answer RQ2.b (Which facets of CPS competence do the assessments measure?), the units of analysis were the assessments, and the information in the articles were all used as the source of data to inform what facets of CPS competence were targeted by the different assessments. The reporting of the content of each assessment was scrutinised with the aim of identifying the facets, i.e., skills within components of CPS competence, measured by the assessments. The CPS framework developed by Oliveri et al. (2017) was revised and applied as a coding template (Table 3.2) to categorise the facets within components targeted by the assessments. To inform revisions, the PISA 2015 CPS framework (OECD, 2017a) and the Framework for teachable CPS skills (Hesse et al., 2015) were examined to accommodate a more specific perspective on assessment across educational settings.

As shown in Table 3.2, the Communication component consisted initially of two skills: ‘Active listening’ and ‘Exchanging information’. Following examination of the frameworks, a third skill, ‘Audience awareness’ was added. This skill refers to the awareness of how to tailor contributions to increase suitability for others (Hesse et al., 2015). The Leadership component initially consisted of five skills. In the revised form, the skill ‘Monitoring performance’ was deemed more relevant to the Problem-solving component, and it was therefore moved and joined with the skill ‘Evaluating solutions’. They both refer to monitoring and reflecting processes relevant to individual problem solving (OECD, 2017a). The skill ‘Transformational leadership’ was already covered in the description of the skill ‘Team empowerment’, and it was therefore dropped from the Leadership component.

Table 3.2. Revised coding framework

Components	Skills	Description
1. Teamwork	Team cohesion	Recognising team members' preferences, strengths, and weaknesses.
	Team empowerment	Being committed to one's team, motivating and inspiring action in others.
	Team learning	Increased knowledge as a result of being a team member.
	Self-management and self-leadership	Participating, monitoring own performance, adjusting own plans, and meeting goals.
	Open-mindedness, adaptability, and flexibility	Incorporating others' ideas and feedback, be open to diverse perspectives, and adapting contributions of others.
2. Communication	Active listening	Giving others the opportunity to speak, interpreting non-verbal cues, and posing follow ups.
	Exchanging information	Communicating with team members to achieve the goals and responding to others.
	Audience awareness	Adapting behaviour and tailoring contributions.
3. Leadership	Organising activities and resources	Managing resources or people and defining roles and responsibilities of team members.
	Reorganising when faced with obstacles	Identifying and correcting gaps or misunderstandings.
	Resolving conflicts	Achieving a resolution of differences or reaching a compromise.
4. Problem Solving	Brainstorming and identifying problems	Exploring and understanding elements of the task and roles to solve the problem.
	Interpreting and analysing information	Identifying and defining tasks to be completed, identifying connections.

	Planning and implementing solutions	Setting goals, enacting plans, implementing solutions, and following rules of engagement.
	Evaluating solutions and monitoring performance	Monitoring results of actions and evaluating success in solving the problem.
	Reaching correct solution	Achieving the desired solution.

Notes. The skills in bold letters represent the revisions/additions to the CPS framework by Oliveri et al. (2017)

The Problem-solving component initially consisted of five skills. In the revised form, the skills ‘Brainstorming’ and ‘Identifying problems’ were joined as one skill relevant to the exploring and understanding process of individual problem solving (OECD, 2017a). Similarly, the skills ‘Planning’ and ‘Implementing solutions’ were joined in a skill relevant to planning and executing process (OECD, 2017a). Finally, the skill ‘Reaching the correct solution’ was added in the Problem-solving component. There are different approaches in dealing with the outcome of a problem-solving task, some researchers give credit to students for arriving at the correct answer (Hesse et al., 2015), while others focus on the process without evaluating the outcome (OECD, 2017a). In assessments composed of multiple problem-solving tasks, reaching the correct solution in one task might influence team cohesion in the next tasks, and therefore, it was added in the revised coding framework.

3.5.5.3 Evaluation of CPS competence measures

To answer RQ2.c (What strategies for validating CPS competence measures are reported?), the units of analysis were the articles, which were coded by extracting information about the strategy for validating CPS measures that they followed. Validity is identified as “the most fundamental consideration in developing tests and evaluating tests” (American Educational Research Association et al., 2014, p. 11). Theoretical validity frameworks (e.g., Kane, 2006; Messick, 1995) and standards for educational testing (American Educational Research Association et al., 2014) are frequently used as a frame of reference in validation studies in the field of educational research.

Differentiating validity into distinguishable validity aspects, helps to highlight issues and nuances that might otherwise be downplayed or overlooked, such as the social consequences of performance assessments (Messick, 1995). As a unified concept, validity integrates considerations of content, criteria, and consequences into a construct framework for empirically testing rational hypotheses about score meaning and utility (Messick, 1995).

Considering validity as a unified concept (Messick, 1995), the six distinct validity aspects emphasising content, substantive, structural, generalisability, external, and consequential aspects of validity, were adopted when coding validity evidence in the reviewed articles. Taken together, these aspects provide a way of addressing the multiple and interrelated validity questions that need to be answered in justifying score interpretation and use.

Table 3.3 provides a description of the coding template used to extract information from the articles undertaking validation work. It explains each validity aspect together with examples of validity evidence.

Table 3.3. Coding template for evaluating validity evidence

Validity aspect	Description of validity aspect	Example evidence
1. Content aspect	Evidence of content relevance, representativeness, and technical quality	Information about item statistical properties/ indicator statistics
2. Substantive aspect	Evidence that the theoretical processes are engaged by respondents in the assessment tasks	Think aloud or retrospective reflections of thought processes for arriving at a response
3. Structural aspect	Evidence about the extent to which the internal structure of the assessment reflected in the scores is consistent with the structure of the construct domain	Sub-scale correlations, dimensionality analysis
4. Generalisability aspect	Evidence about the extent to which score properties and interpretations generalise appropriately to and across population groups, settings, and tasks	Reliability indicators (e.g., Cronbach's alpha)
5. External aspect	Evidence about the relationship of test scores to variables external to the test, hypothesised to measure the same construct, related or different constructs (convergent and discriminant evidence)	Theoretically predicted association between the test score and other variables
6. Consequential aspect	Appraisal of the implications of score interpretation as a basis for action	Appraisal of potential and actual social consequences of the applied testing

Notes: Table is own adaption using information provided by Messick (1995) and American Educational Research Association et al. (2014).

3.6 Methodology for Rasch analysis of PISA 2015 CPS measure and its correlates with important variables (Chapter 6)

3.6.1 Overview

In the first empirical phase (Chapter 6) the validity evidence for the PISA 2015 CPS competence measure was explored by analysing quantitative data. This phase aims to contribute to the validation of the PISA 2015 CPS competence measure using a definition of validity as a unified concept (Messick, 1995) and applying the Rasch measurement framework for the scaling of items (Rasch, 1960). Specifically, secondary data analysis in this phase involved a sequential two-step procedure including (i) measure validation and (ii) modelling with constructed measures.

3.6.2 Data

The analysis presented in Chapter 6 was based on data from the 2015 cycle of the Programme for International Student Assessment (PISA), led by the Organisation for Economic Co-operation and Development (OECD). The dataset is publicly available for secondary analysis at the official website[4] of the PISA study. PISA is an international comparative survey study, which is conducted every three years, starting from 2000, and provides evidence on how the achievement of 15-year-olds in schools varies across countries. As an age-based survey, the age 15 is specifically selected since these students are approaching the end of compulsory schooling in most participating countries. One of the main aims of the PISA study, which differentiates it from other large-scale assessments such as the Trends in International Mathematics and Science Study (TIMSS), is to evaluate the extent to which students are prepared to meet the challenges of today’s knowledge societies and can apply the knowledge and skills they have learned and practised at school in unfamiliar settings (OECD, 2017a). Taking what is also called a literacy perspective (Turner & Adams, 2007), PISA does not strive to assess curricula directly, but rather assess students’ ability to apply knowledge in new situations.

[4] https://www.oecd.org/pisa/data/

In each PISA cycle, students are tested in three core subjects, science, mathematics, reading, with one of these subjects being the particular focus, also called major domain, in each cycle, and thus covered in more detail (e.g., including the assessment of student attitudes related to the subject with major focus). The PISA 2015 survey focused on science, with reading and mathematics as minor areas of assessment. Additionally, in most PISA cycles, cognitive assessments of an "innovative domain" have been included, which are basically assessments of additional cross-curricular competencies (OECD, 2017a). For PISA 2015, this domain was collaborative problem solving, while in 2012 it was individual problem solving.

For the first time in PISA 2015, the main mode of assessment for all subjects was computer-based tests[5], which lasted a total of two hours for each student, and covered science, mathematics, reading, and CPS (OECD, 2017c). Finally, students answered a (background) questionnaire, which took around 30 minutes to complete and sought information about various aspects of their home, family, and school background as well as their attitudes and learning experiences. Since CPS was added as an innovative domain in the PISA 2015 cycle, the items developed were not administered again in other cycles, and therefore data from that administration were used to answer the research questions.

3.6.3 Sampling

The PISA 2015 target population in each country and economy is students of a specific age[6] attending educational institutions. Specifically, students are aged between 15 years 3 months and 16 years 2 months at the time of assessment (OECD, 2017a). The PISA 2015 study was conducted between November and December 2015. A two-stage sample design was used for the selection of schools

[5] Paper-based assessment instruments for science, mathematics, and reading were provided as alternative to computer-based testing to countries and economies that did not have the resources available in schools (OECD, 2017c).

[6] It has been argued that school grade levels are often not good indicators of student cognitive development and therefore not preferred as a criterion for sample selection (OECD, 2017a).

and students to take part in the study. The first stage involved the selection of schools (the primary sampling unit), with the probability of selection being proportional to the size of the school, but other variables were also used depending on the country. For England, schools were randomly selected to be representative of the national distributions of school type (e.g., independent, academy), location and historical GCSE performance (Department for Education, 2017). Within each school, students were then randomly selected (n = 30 students). Approximately 540,000 students in schools from 72 different countries and economies took part in PISA 2015 (OECD, 2017b).

The PISA 2015 data for England are representative of the target population[7] (Department for Education, 2017). The analysis presented in Chapter 6 used data from PISA 2015 for England, where a total of 5,194 students in 206 schools (2,475 female, 2,719 male) took part. From those, a total of 1,584 students took the PISA 2015 CPS assessment (more details are presented in the following section 3.6.4 PISA 2015 test design) and were therefore the analytical sample of the corresponding paper. Further details on the technical standards can be found in the PISA 2015 technical report (OECD, 2017c) and the National Reports produced (Department for Education, 2017; Jerrim & Shure, 2016).

3.6.4 PISA 2015 test design

The PISA study employs a complex test design within the space of a 2-hour test. In PISA 2015 cycle, the study included 184 questions in science (major domain of assessment), 81 questions in mathematics (minor domain), 103 questions in reading (minor domain), and 117 questions in CPS (innovative domain). Due to time constraints, participating students did not answer the entire set of questions developed for every subject. Instead, and to maximise the allocated time, questions

[7] "Although the PISA 2015 data for England are representative of the target population, the fact they are based upon a sample (rather than a census) means there will be a degree of uncertainty in all estimates derived using these data" (Department for Education, 2017, p. 16).

were divided into several subject-specific clusters, taking about 30 minutes to complete, which were then organised in different test forms (also known as booklets).

Each test form consisted of four different clusters. Individual students were then randomly allocated one of these test forms to complete, and therefore, they undertook a sub-set of the entire assessment material available. Each of the subject-specific cluster appears in different position in the test forms, and the cluster allocation is carried out in a way that ensures cluster overlap from one cluster grouping to another (OECD, 2017c). Appendix 3 presents the total of 66 different test forms created for PISA 2015 computer-based delivery and the respective sample sizes for England. All students answered two clusters of science questions (major domain of assessment), which approximates 1 hour of test time. For the minor and innovative domains of assessment, only around 40% of students answered any questions in reading, 40% of students answered any questions in mathematics, and 30% of students answered any questions in CPS (OECD, 2017c, p. 40). For England, a total of 1,584 students took the CPS assessment.

3.6.5 PISA 2015 Collaborative problem-solving competence assessment

The PISA 2015 CPS assessment was designed to capture the competence of individuals to work in collaborative settings, which was achieved by having students interact with pre-programmed computer-simulated partners (also known as computer agents) instead of other humans. Across different problem scenarios, computer-simulated partners were programmed to match different roles, attitudes, and levels of competence to vary the CPS situation.

This approach has been argued to allow the high degree of control and standardisation required for large-scale international assessment as well as measurement in multiple situations within the time constraints of the PISA test (OECD, 2017a). Each student completed the PISA 2015 CPS assessment on a computer individually. The measurement focused on the outputs of the individual, rather than outputs from the rest of the group or the overall performance of the group, and analysis was performed at the student level.

Questions included in the CPS domain (i.e., CPS items) were organised in different assessment tasks[8]. These were interactive problem-solving scenarios, based on a common piece of stimulus, that students had to work through. Six assessment tasks, ranging from 12 to 28 items, were organised in subject-specific clusters of CPS as shown in Table 3.4. Students had to respond to items in the tasks by performing an action, which could be any explicit act made that could change the state of the collaborative problem (OECD, 2017b).

Table 3.4. Assessment tasks and clusters of CPS in PISA 2015

Cluster of CPS	Assessment task	Number of items
CPS1	Meeting in the park	15
	Making a film	23
CPS2	Field trip	12
	Preparing a presentation	27
CPS3	Xandar	12
	The garden	28

Specifically, students could either make a multiple-choice selection of predefined messages presented to them in a chat area to communicate with their team members (see an example item in the following section 3.6.6), or perform actions (e.g., dragging and dropping) in the visual display area (OECD, 2017b). No free-response items were available to students. In addition, items were independent of one another, meaning that irrespective of which response a student had selected for a particular item, the computer-simulated partners responded in a way that all

[8] The original term used in PISA's nomenclature to describe assessment tasks was "units": each unit contains one problem scenario and several items (OECD, 2017a, p. 151). To maintain consistency across the research papers in the thesis, and across the various literatures that use the terms "task" and "unit" interchangeably to describe a set of items organised in a problem-solving scenario, the term "task" is adopted throughout this thesis.

students were faced with an identical version of the next item. Appendix 4 gives details about scoring and targeting for the entire set of items included in the PISA 2015 CPS dataset.

Table 3.5 presents the distribution of items by assessment task and CPS skill. A total of 117 items were included in the PISA 2015 CPS assessment. Each item as a unit of measurement targeted one of the 12 CPS skills identified in the PISA 2015 CPS framework (OECD, 2017a), however CPS skills were not equally represented in the assessment (Table 3.5). The CPS dataset consisted of 97 items coded as dichotomous and 20 items coded as polytomous. Depending on students' actions, they received full credit, partial credit, or no credit, which refer to correct, partially correct, and incorrect response categories respectively. Overall, PISA 2015 measured students' performance in CPS on a single scale that provided an overall assessment of 15-year-olds' CPS competence (OECD, 2017c).

Table 3.5. Distribution of PISA 2015 CPS items by assessment tasks and CPS skills

Collaborative problem-solving competencies	Collaborative problem-solving skills	Meeting in the Park	Making a film	Field trip	Preparing a presentation	Xandar	The garden	Total
Establishing and maintaining shared understanding	**A1.** Discovering perspectives and abilities of team members	1	8	-	8	1	2	20
	B1. Building a shared representation and negotiating the meaning of the problem	5	-	2	7	2	8	24
	C1. Communicating with team members about the actions to be/being performed	2	-	-	-	1	2	5
	D1. Monitoring and repairing the shared understanding	-	2	1	5	1	3	12
Taking appropriate action to solve the problem	**A2.** Discovering the type of collaborative interaction to solve the problem, along with goals	1	1	-	-	-	-	2
	B2. Identifying and describing tasks to be completed	2	2	1	-	-	-	5
	C2. Enacting plans	1	2	3	5	-	5	16
	D2. Monitoring results of actions and evaluating success in solving the problem	1	-	1	-	1	-	3
Establishing and maintaining team organisation	**A3.** Understanding roles to solve the problem	-	-	-	-	-	-	0
	B3. Describing roles and team organisation	1	2	-	-	3	2	8
	C3. Following rules of engagement	-	5	4	2	2	1	14
	D3. Monitoring, providing feedback and adapting the team organisation	1	1	-	-	1	5	8

Notes: Table is own adaptation using information from the PISA 2015 CPS framework (OECD, 2017a) and PISA 2015 Technical report (OECD, 2017c).

To help interpret PISA scores, the derived PISA 2015 CPS scale was split in five levels of proficiency as presented in the Table 3.6 below. Levels 1 to 4 are described based on the skills needed to successfully complete the items located within those levels, while the last level (below level 1) is defined based on the absence of these skills (OECD, n.d.). Level 1 is the lowest level corresponding to an elementary level of CPS skills and Level 4 is the highest described level.

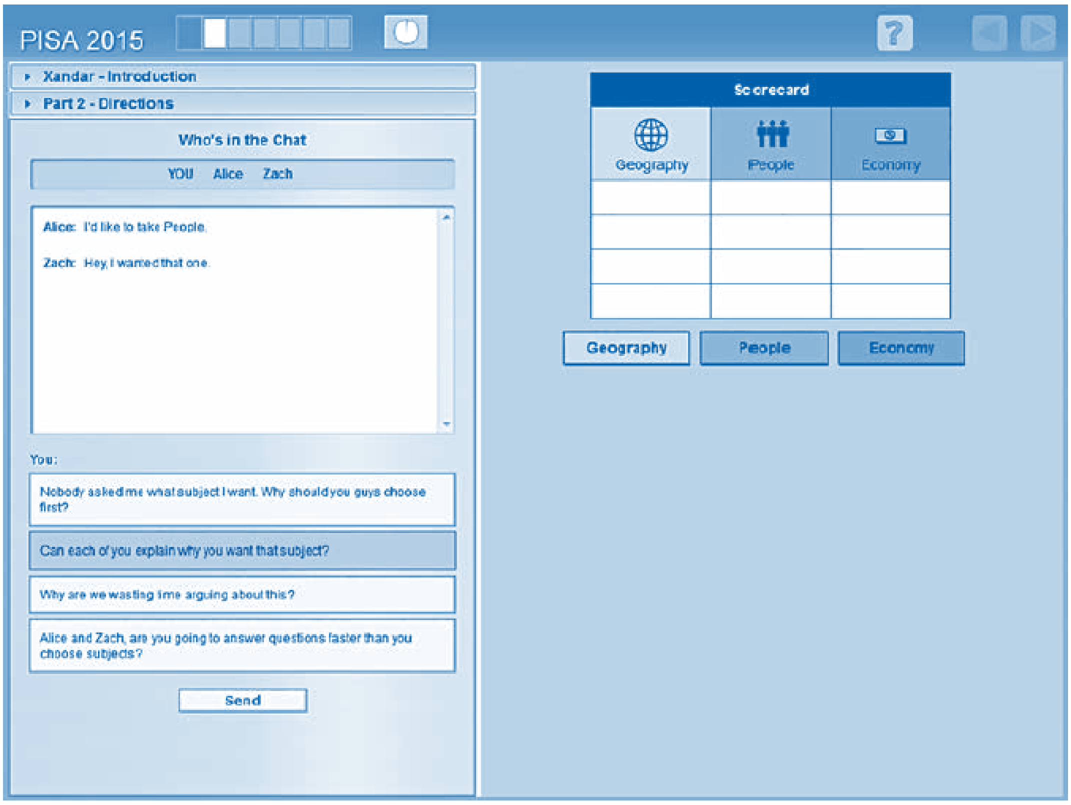
Table 3.6. PISA 2015 - Levels of proficiency in collaborative problem solving

Level	**What students can typically do**
Level 4	Students can successfully carry out complicated problem-solving tasks with high collaboration complexity. They take initiative and perform actions or make requests to overcome obstacles and to resolve disagreements and conflicts. They can balance the collaboration and problem-solving aspects of a presented task, identify efficient pathways to a solution, and take actions to solve the given problem.
Level 3	Students can complete tasks with either complex problem-solving requirements or complex collaboration demands. They can recognise the information needed to solve a problem, request it from the appropriate team member, and identify when the provided information is incorrect. When conflicts arise, they can help team members negotiate a solution.
Level 2	Students can contribute to a collaborative effort to solve a problem of medium difficulty. They can help the team establish a shared understanding of the steps required to solve a problem. These students can request additional information required to solve a problem and solicit agreement or confirmation from team members about the approach to be taken.
Level 1	Students can complete tasks with low problem complexity and limited collaboration complexity. They can confirm actions or proposals made by others. They tend to focus on their individual role within the group. With support from team members, and when working on a simple problem, these students can help find a solution to the given problem.
Below level 1	Absence of skills described in Level 1.

Notes: Information adapted from PISA 2015 Results report (OECD, 2017b, p. 74)

3.6.6 An example CPS item

An example item from the only publicly available CPS task[9] is briefly reviewed. In the CPS task Xandar, a three-person team consisting of the student test-taker and two computer-simulated partners was asked to take part in a contest where they had to answer questions about the fictional country of Xandar (OECD, 2017b). The sample item (item 83) illustrated in Figure 3.5 required students to help team members negotiate a solution when conflict arises (OECD, 2017b). In this case, both team members (Alice and Zach) wanted to answer questions from the same subject area. The credited response to this item was the message: “Can each of you explain why you want that subject?”. This was expected to solicit additional information about each team member’s point of view (OECD, 2017b). This item reflected the CPS skill “Discovering perspectives and abilities of team members”.



Alice: I'd like to take People.

Zach: Hey, I wanted that one.

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

Can each of you explain why you want that subject?

Why are we wasting time arguing about this?

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

Figure 3.5. Screenshot of a released PISA CPS item from the Xandar task (OECD, 2017b).

[9] Available at: http://www.oecd.org/pisa/test/

Notes: Chat space displays the pre-defined messages for communication with the computer-simulated agents, and task space is where actions are performed. Second message is the credited response. Reported item difficulty level is Level 3. Material used under Creative Commons Attribution-NonCommercial-ShareAlike 3.0 IGO (CC BY-NC-SA 3.0 IGO) license.

3.6.7 Analytical approach

Analysis involves a sequential two-step procedure: the construct validation and the modelling with constructed measures. It should be noted that in both steps of the analysis Messick's (1989) definition of validity as a unified concept has been adopted. Specifically, validation is considered as a continuing process referring to the accumulation of evidence to support validity arguments. Following Messick, validity is defined as "an overall evaluative judgment of the degree to which evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores and other modes of assessment" (Messick, 1995, p. 741). Six distinguishable validity aspects are adopted to emphasise content, substantive, structural, generalisability, external, and consequential aspects of construct validity (Messick, 1989, 1995).

In **step one**, students' responses to the PISA 2015 CPS assessment are used and a series of Rasch measurement models are employed. First, the Partial Credit model is used to analyse all CPS items, assuming that they measure a single construct (i.e., CPS competence), ignoring any distinctions between possible underlying dimensions/sub-scales. Second, the Partial Credit model is used to analyse separately the three hypothesised dimensions that are based on the different CPS competencies that the items target. The validation process is conducted within the Rasch measurement framework and follows widely accepted Rasch guidelines (Wolfe & Smith, 2007a, 2007b), which are in turn based on Messick's (1989) definition of validity.

In **step two**, the constructed CPS competence measures are used as variables in further statistical analyses, including regression modelling, to evaluate external and consequential aspects of validity (Wolfe & Smith, 2007b). Specifically, correlations between sets of measures are used to report similarities between related constructs (i.e., attitudes towards collaboration and CPS competence measures). Regression modelling is also utilised to determine whether (theory-based) predictions about changes within individuals are realised in the measures of CPS competence.

3.6.8 Construct validation

3.6.8.1 Scaling and Item response theory

Classical test theory was the dominating approach to testing for much of the middle part of the 20th century across the world, although it has been argued to suffer from several limitations (Berezner & Adams, 2017; Hambleton et al., 1991; Panayides et al., 2010). Since classical test theory does not make any assumptions about latent variables, inferences beyond the set of items being tested cannot be made (Wu & Adams, 2007). Another drawback is that the item statistics (difficulty, discrimination, reliability) are examinee dependent, and no information is available about how examinees of specific abilities might perform on a certain test item (Hambleton et al., 1991; Panayides et al., 2010).

Item response theory (IRT) is one of the approaches developed to deal with the limitations of classical test theory and is a commonly used technique for scaling items (Hambleton et al., 1991). The process of scaling, in which raw data are converted to numerical indicators, allows both measurement of an underlying construct, and comparison between sets of items (Berezner & Adams, 2017). The main idea of IRT is to use a mathematical model for predicting the probability of success of a person on an item, depending on the person's "ability" and the item "difficulty" (Wu & Adams, 2007, p. 13). As described by Panayides et al. (2010, p. 616), it provides alternative models with the following desirable features: item characteristics are not group dependent, scores describing examinees' abilities are not test dependent, a measure of precision for each ability score is produced, and the probability that an examinee of any ability will answer items of any difficulty correctly is estimated.

IRT models are mathematical models that relate observed categorical variables (responses to test items) to hypothesised unobservable latent variables (proficiency in a subject) (Berezner & Adams, 2017). The latent variables are the constructs to be measured, and the term 'latent' is used to emphasise that they are not directly observable. For that reason, items are used to tap into the latent variable, with a person's responses to items being observable. As illustrated by Wu and Adams (2007), the items represent little ideas based on the bigger idea of the latent variable. For example, if the latent variable is CPS competence, then the items are individual questions about specific skills in CPS. Through a person's item response patterns, inferences about a person's level on the latent variable can be made (Wu & Adams, 2007). By modelling the relationship between raw data and an assumed underlying construct using IRT models, interval scale scores can be produced (Berezner & Adams, 2017).

Under IRT, it is assumed that each student being assessed can be characterised as having or holding some amount of an underlying construct that will influence how well that student will perform on an assessment targeting that construct. This amount is a quantity on a continuous metric and it is referred to as a 'person parameter' (Berezner & Adams, 2017, p. 325). Taking CPS competence as the underlying construct of interest, individuals possessing greater CPS competence are expected to perform better on an assessment of that construct than individuals with less CPS competence.

3.6.8.2 Item response theory models used in PISA 2015

In prior PISA cycles (2000-2012), the Rasch model (1960) and the Partial Credit model (Masters, 1982) were used to estimate item difficulty parameters, i.e., calibrate/scale the items (OECD, 2017c). Concerns were raised by some researchers over the insufficiencies of the Rasch model to adequately address the complexity of the PISA data (e.g., Kreiner & Christensen, 2014), which led to technical changes in PISA 2015. To address these concerns, PISA 2015 implemented for the first time the two-parameter-logistic model (Birnbaum, 1968) for dichotomously scored responses and the generalised Partial Credit model (Muraki, 1992) for items with more than two ordered response categories (OECD, 2017c, p. 142).

3.6.8.3 The Rasch model

The Rasch model (Rasch, 1960), is a mathematical model that was developed during the 1960's by a Danish mathematician named Georg Rasch (Wright & Stone, 1999). It is also referred to as a one-parameter item-response model, since a single parameter, i.e., item difficulty, is used to describe each item. The model proposes a mathematical relationship between a person's 'ability', the difficulty of the item, and the probability of the person answering correctly that item (Wright, 1999; Wright & Mok, 2000).

The principle of the Rasch model is the following:

> "a person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one" (Rasch, 1960, p. 117).

The Rasch model is used for dichotomous items, when only two possible response categories are defined, usually correct (1) and incorrect (0). It performs a logarithmic transformation of the items and person data to transform ordinal into interval data, which yields one scale expressed in 'logits' (log odds units), for both person ability and item difficulty (Bond & Fox, 2007). The probability of a correct response is modelled as a logistic function of the difference between the ability of the person and difficulty of the item. The construction of a logit scale (based on the Rash model) is given through the following equation:

$$\log \left(\frac{Probability\ of\ success}{Probability\ of\ failure}\right) = \textit{Ability - Difficulty}$$

To compute the probability of achieving a score of 1 (correct response) rather than 0 (incorrect response) on item i, the Rasch model is expressed in the following form:

$$P(X_i = 1|\theta,\delta_i) = \frac{\exp(\theta - \delta_i)}{1 + \exp(\theta - \delta_i)}$$

where θ is the 'ability' of the person and δ*i* the difficulty of item i (Bond & Fox, 2007). These two parameters (difficulty and ability) estimates are mapped on the same (interval) scale. The logit scale that is created is independent of the particular set of items or the particular sample of persons that have been used to calibrate the items (Wright, 1999). If person 'ability' matches the item difficulty, the expected probability of success on the item is equal to 0.5. In other words, item difficulty under the Rasch model can be interpreted as the location along the 'ability' continuum at which a person is just as likely to answer the item correctly or incorrectly (OECD, 2017c). It follows that, if the item difficulty is higher than the person 'ability', the probability of getting a correct response will be lower, and if the item difficulty is lower than the person 'ability', then the probability of getting the item right will be higher.

One of the advantages of the Rasch model, is that the single scale enables both persons and items to be placed on the same continuum defining an underlying variable (Griffin et al., 2015). Finally, a special property of the Rasch model is specific objectivity, that it, the comparison between two persons should not be influenced by the specific items used for the comparison (Wu & Adams, 2007). Similarly, the comparison between two items should not be influenced by which person took the two items. This sample-free property of the Rasch model is considered important, since it allows, for example, making statements about relative item difficulties without reference to specific persons (Wu & Adams, 2007).

3.6.8.4 Assumptions of the Rasch model

The following key assumptions are at the core of the Rasch model and are important to be considered in the validation process.

Unidimensionality: It involves the key idea that there is a single underlying construct (latent variable/trait) being measured. As stated by Wright and Linacre (1989), unidimensionality is a qualitative rather than quantitative concept, in the sense that no actual test can be perfectly unidimensional. All data are multidimensional to some extent (Panayides et al., 2015). Many psychometricians (e.g., Masters & Keeves, 1999; Smith, 2002; Wright & Linacre, 1989) have suggested that unidimensionality does not implicitly mean only one dimension, but rather the presence of a dominant dimension and possibly of minor dimensions (Panayides et al., 2010). Extra dimensions may reflect different person response styles, different item content areas or be an artefact of test construction (Panayides et al., 2010).

Local independence: A student's response to one item should not affect the student's responses to other items. This assumption is violated when the probability of success on an item depends on the response on another item (Wu & Adams, 2007). For example, an item provides clues to the answer of another item. If two items are locally independent, then success or failure on one item does not affect the probability of succeeding on the other item.

Item discrimination: It refers to the power of the items to discriminate between the more and less able respondents. This assumption indicates the extent to which success on an item corresponds to success on the whole test (Kelley et al., 2002).

3.6.8.5 The Partial Credit model

As has been already stated, the Rasch model is used only for dichotomously scored items. However, some items in PISA 2015 CPS assessment were scored polytomously, i.e., more than two ordered response categories were possible. Therefore, an extension of the Rasch model appropriate for items scored polytomously, the Partial Credit model, is used (Masters, 1982; Wright & Masters, 1982). This model incorporates the possibility of having one or more intermediate levels of success (i.e., partially correct answers) for different items on the same test (Bond & Fox, 2007). For this reason, it is highly applicable in educational testing situations in which partially correct marks are awarded for partially correct answers (Bond & Fox, 2007). The Partial Credit model is expressed in the following form:

$$\Pr\{X_{ni} = x\} = \frac{\exp\left(x(\beta_n - \delta_i) - \sum_{k=1}^{x} \tau_{ki} \right)}{\sum_{x=0}^{m_i} \exp\left(x(\beta_n - \delta_i) - \sum_{k=1}^{x} \tau_{ki} \right)}$$

where $x \in \{0,1,2,...,m_i\}$ is the integer response variable for person *n* with ability β_n responding to item *i* with difficulty δ_i, and $\{\tau_{1i}, \tau_{2i}, ..., \tau_{mi}\}$ are thresholds between the $m_i + 1$ ordered categories (Pampaka et al., 2013, p. 202).

As in the Rasch model, greater ‘ability’ corresponds to a higher probability of achieving a larger score on an item. The model can be applied to any set of test data collected for the purposes of measuring achievements or attitudes, given that responses to each item are scored in two or more ordered categories (Masters, 1999). It follows that, the Partial Credit model is appropriate for the validation of the measure under investigation due to the nature of student responses.

First, the Partial Credit model is used to analyse all 117 items, assuming a single construct, ignoring any distinctions between possible underlying dimensions/sub-scales. Second, the Partial Credit model is used to analyse separately the three hypothesised dimensions that are based on the different CPS competencies that the items target, based on the PISA CPS framework. These are “Establishing and maintaining shared understanding”, “Taking appropriate action to solve the problem”, and “Establishing and maintaining team organisation”. Each hypothesised dimension is modelled as a unidimensional construct producing three separate student estimates. The WINSTEPS software (Linacre, 2006c) is used to perform this analysis.

3.6.8.6 More complex item response theory models and the advantages of Rasch measurement

The historical use of the Rasch model in the PISA study (from PISA 2000 to 2012) has received criticism over the years (e.g., Fernandez-Cano, 2016; Goldstein, 2017; Kreiner & Christensen, 2014). As has been already stated, in PISA 2015 the two-parameter logistic model has been used to address concerns raised about usage of the Rasch model. At the same time, other international studies such as TIMSS utilised more complex IRT models such as the three-parameter IRT model. As explained by Jerrim et al. (2018), the essential difference between the two-parameter logistic model and the Rasch model is that, in the former model, questions are not only allowed to vary in terms of their difficulty, but also their discrimination (i.e., how well each item in PISA 2015 is thought to measure students' reading/ science/mathematics/CPS skills). Nevertheless, little evidence was found to support that moving to a more complex model such as the two-parameter-logistic model had any meaningful impact upon cross-country comparisons of the PISA 2015 results (Jerrim, Parker, et al., 2018). It was also argued that some of the media reports questioning the historical use of the Rasch model in the PISA study have been overblown (Jerrim, Parker, et al., 2018).

When discrimination is used as additional parameter, as it happens in the case of the two-parameter logistic model, the simple one-to-one relationship with the raw score is lost (Griffin et al., 2015). Additionally, it has been argued that, when comparing the two-parameter logistic model with the Rasch model, it is important to distinguish between measurement and modelling (Panayides et al., 2010). Specifically, the Rasch model corresponds to the principles of measurement, whereas other more complex item response theory models correspond to modelling. In other words, the two-parameter logistic model seeks to fit a model to the data, while the aim of Rasch measurement approach is to measure and not to accommodate the data. Therefore, if the purpose of the researcher is to construct a good measure, then the items and the test should be constrained to and valuated by the principles of measurement (Panayides et al., 2010).

Considering all the above, the family of the Rasch model has been chosen for the analysis presented in Chapter 6. Among the advantages of the Rasch measurement are that it can produce linear measures, overcome missing data, give estimates of precision, have devices of detecting misfit, and map both items and respondents on the same 'logit' scale (Wright & Mok, 2000, 2004). It has been argued that only the family of Rasch measurement models has these characteristics (Panayides et al., 2010) and that fundamental measurement in the social sciences is obtainable only through the Rasch measurement (Wright, 1983). In general, the Rasch model is chosen for construct validation, since it allows establishing the relative difficulty of each item in recording students' CPS competence, from the lowest to the highest levels the instrument is able to record (Bond & Fox, 2007). The creation of a single scale linking a person's 'ability' and item difficulty also entails practical advantages in teaching. For example, once a student has been located on the scale, features of items at about the level of the student's ability can be examined to draw inferences about the kinds of items they are likely to be able to complete and suggest a focus for future teaching for that student (Ramalingam, 2016).

When constructing measures of CPS competence, other researchers (e.g., Griffin et al., 2015; Harding et al., 2017) have also preferred the Rasch model over more complex IRT models for interpretative purposes. The Rasch model is considered appropriate for the aims of this phase, i.e., construct validation, since it can be used to evaluate a set of items as a social science measure and construct new variables. In addition, assuming the data fit the model, then the interval scores (logits) of student CPS competence (or any other construct) can be used in further analysis. Finally, the Rasch model provides a set of guidelines that have been widely used for measure construction and validation (Pampaka, 2021; Pampaka et al., 2013; Wolfe & Smith, 2007a, 2007b), which are also adopted in the current work and are described in the following section.

3.6.8.7 Validation using the Rasch model

Decisions about validity are informed by different statistical indices, such as item fit statistics, dimensionality diagnostics, reliability and separation statistics, person-item maps, and differential item functioning (Bond & Fox, 2007). These are discussed in more detail in this section.

Item fit statistics

In the Rasch measurement context, traditional item fit statistics, i.e., standardised (Zstd) and mean-square (MNSQ) fit statistics, indicate how accurately the data fit the model, and this, as a common practice, provides evidence in support (or not) of the unidimensionality assumption. Fit statistics can therefore be used as indicators of validity (Wolfe & Smith, 2007b). Inconsistent data (e.g., misfit items or persons) may become a source of further inquiry, and they may suggest the possibility of existence of new dimensions in the data, hence lack of unidimensionality (Pampaka et al., 2013).

Mean-square (MNSQ) weighted fit statistic (also referred to as INFIT) and unweighted fit statistic (also referred to as OUTFIT) are examined as evidence that the underpinning construct is represented by the items, and possible misfit items are identified. Statistically, mean-squares are chi-square statistics divided by their degrees of freedom and their expected values are close to 1.0. Values substantially less than 1.0 may indicate observations are too predictable (redundancy, data overfit the model) and values greater than 1.0 may indicate unpredictability (unmodeled noise, data underfit the model). A range of 0.70 and 1.30 is often used to describe acceptable fit (Adams & Khoo, 1995; Wright & Mok, 2000). However, it is important to note that the issue of cut-off points to these values still remains unresolved in the Rasch measurement literature (Pampaka et al., 2013) and most of the practitioners use them as guides for flagging concerning issues depending on the context and stakes of the study, rather than strict decision making tools (e.g., Bailey et al., 2017; Grant et al., 2019; Prevett et al., 2020; Whelehan et al., 2021). For this thesis, I consider existing guidelines and previous research (Pampaka, 2021;

Pampaka et al., 2013) and I consider values above 1.3 to suggest causes for concern warranting discussion and explanation.

Standardised fit statistics (Zstd) are t-tests of the hypothesis that data fit the model perfectly. These report the statistical significance (probability) of the chi-square (mean-square) statistics occurring by chance when the data fit the Rasch model (Linacre, n.d.-a). A Zstd value should be flagged as significant if the absolute value is larger than 1.96. Residual correlations for items are also examined to check for items that may be locally dependent. High positive residual correlations (around 0.70) may indicate local item dependency between pairs of items (Linacre, n.d.-b). Two items may be locally dependent when they duplicate similar features.

Principal components analysis of residuals

Evidence relating to the structural aspect of validity can be explored through the results of a principal components analysis of the residuals, after the Rasch model was fitted, which can aid in determining whether the measure under investigation approximates a unidimensional measure (Linacre, 1998; Smith, 2002; Wolfe & Smith, 2007b). The WINSTEPS software, used for this analysis, places eigenvalues from the Rasch-scaled measures and the principal component analysis of the residuals onto a common scale (Wolfe & Smith, 2007b). Observed patterns in the residuals can indicate multidimensionality (Smith, 2002). Eigenvalues are examined to determine "whether there is sufficient amount of variance accounted for by components beyond that accounted for by the Rasch measures to justify further exploration of the dimensionality of the measures" (Wolfe & Smith, 2007b, p. 213).

If the eigenvalue of the first contrast (i.e., the component that explains the largest possible amount of variance in the residuals) is small, usually less than two units, then the first contrast is at the noise level (Linacre, 2006a). If not, the loadings on the first contrast indicate that there are contrasting patterns in the residuals, which could be considered as evidence of multidimensionality, existence of sub-scales, or at least as a concern for violation of unidimensionality. When analysis of measures that were intended to be unidimensional indicates that those measures are not adequately described by a single dimension, then further analysis/investigation of the dimensionality of the measures is warranted (Wolfe & Smith, 2007b).

Reliability and separation

The person and item separation statistics in Rasch measurement provide an analytical tool by which to evaluate the successful development of a measure and with which to monitor its continuing utility (Wright & Stone, 1999). Person separation is used to classify people (in this case students), whilst item separation is used to verify the items' hierarchy (Pampaka, 2021). Low person separation (lower than 2) implies that the instrument may not be sensitive enough to distinguish between high and low performers, and therefore, more items may be needed (Linacre, 2006b). Item separation indicates how well a sample of people can separate those items used in the test (Wright & Stone, 1999). Low item separation (lower than 3) implies that the person sample is not large enough to precisely locate the items on the latent variable, i.e., to confirm the item difficulty hierarchy of the instrument (Linacre, 2006b). Where these statistics are expressed as reliabilities, they range from 0.0 to 1.0, the higher the value the better the separation that exists and the more precise the measurement (Wright & Stone, 1999).

Person-item maps

In addition to item fit statistics, an important step in validating a scale is to assess targeting. Person-item maps and the item difficulty hierarchies provide evidence for substantive, content, and external aspects of validity (Wolfe & Smith, 2007b). A person-item map (also called Wright map) visualises the location of the item difficulties and the distribution of respondents' 'abilities'. 'Ability' is used here as a technical term to refer to the construct under study, in this case, students' CPS competence. The person-item map presents both respondents and items along the same construct (latent trait) describing students' CPS competence (represented by the vertical axis). The left side of the person-item map shows the estimated latent 'ability' distribution, and the right side of the person-item map shows the item hierarchy based on their difficulty.

The map allows direct graphical comparisons between respondents, between items, and between respondents and items (Cascella & Pampaka, 2020). It is ideal that the item difficulty distribution will cover the span of the student 'ability' distribution, thus providing accurate measures of student proficiency over the whole scale (Liu et al., 2008). However, if the distribution of student 'ability' is skewed as compared to the item difficulty distribution, then more items might be needed to capture appropriately the construct.

Item characteristic curve plots

The probability of success on an item for respondents with varying ability is plotted as an item characteristic curve (Wu & Adams, 2007). Item characteristic curve plots visualise the relationship between student latent 'ability' and the probability of a response to a given category within an item. For example, in a dichotomous item (0, 1) a student with high 'ability' shows a probability of success close to 1, a student with low 'ability' shows a probability of success close to 0, and a student with average 'ability' shows a probability of success close to 0.5 (Wu & Adams, 2007).

Establishing measurement invariance – differential item functioning analysis

Evidence relating to the generalisability validity aspect often takes the form of differential item functioning (DIF) analysis, which seeks to determine whether two groups have different probabilities of providing a particular response to individual items, when matched on measures of the construct (Wolfe & Smith, 2007b). DIF suggests group differentiation of the construct measures, an important aspect of validity when an instrument is used with different groups (Pampaka et al., 2013; Thissen et al., 1993). Items are intended to function invariantly with respect to irrelevant aspects of the respondents, i.e., personal features such as gender. Relevant guidelines (Zwick, 2012), for DIF size and its statistical significance are considered for the evaluation of its effect ($|DIF| < 0.43$ is negligible, $0.43 \leq |DIF| < 0.64$ is slight to moderate, and $|DIF| \geq 0.64$ is moderate to large). It is important to acknowledge that it is challenging to disentangle DIF from potential bias, and this involves a combination of conceptual and statistical evidence (Pampaka, 2021). One way, proposed to determine whether group differences are real in relation to the intended construct and its uses, is differential bundle functioning (Ong et al., 2011, 2013). This approach was proposed as a way to build conceptual modelling into the validation of the items (Pampaka, 2021).

3.6.9 Statistical modelling with constructed measures

Following construct validation, the derived scores (i.e., students' measures) are added to the original dataset along with the other student variables for further modelling. Students' CPS competence, their attitudes towards collaboration, performance in science, mathematics, and reading as well as various background characteristics, were the variables used for further analyses.

3.6.9.1 Variables

CPS competence measures: overall and sub-scales

The resulting person (student) scores of the four constructed CPS competence measures (overall score and three sub-scale scores), are the main outcome variables used.

Attitudes towards collaboration

As part of the PISA 2015 background questionnaire, students were asked about their attitudes towards collaboration resulting in two variables: valuing relationships and valuing teamwork. Valuing relationships, is related to altruistic interactions, when the student engages in collaborative activities and valuing teamwork is related to what teamwork can produce, as opposed to working alone (OECD, 2017c). Items presented in Table 3.7 were scored so as higher values correspond to more positive attitudes towards collaboration. The two scales' reliabilities (Cronbach's Alpha) were 0.723 and 0.821[10], respectively (OECD, 2017c, p. 307).

[10] Scale reliabilities correspond to sample from the United Kingdom as reported in the OECD technical report.

Table 3.7. Attitudes towards collaboration items

Variable name	Variable name in PISA 2015	Items	Response categories
Valuing relationships	COOPERATE	I am a good listener. I enjoy seeing my classmates be successful. I take into account what others are interested in. I enjoy considering different perspectives.	“strongly disagree”, “disagree”, “agree”, and “strongly agree”
Valuing teamwork	CPSVALUE	I prefer working as part of a team to working alone. I find that teams make better decisions than individuals. I find that teamwork raises my own efficiency. I enjoy cooperating with peers.	“strongly disagree”, “disagree”, “agree”, and “strongly agree”

Notes: Adapted from OECD (2017c).

Science, mathematics, and reading performance

Due to PISA’s rotated test design, not all participating students answered every test question in every subject. Consequently, instead of generating a single achievement estimate per pupil, a set of ten plausible values were drawn for each pupil in each subject area tested in PISA (Jerrim et al., 2019). These have a mean of around 500 points and a standard deviation of around 100 points. I used the ten plausible values included in the PISA 2015 dataset for each subject (i.e., science, mathematics, reading). Following recommended practice (OECD, 2009), each model was estimated ten times (once for each plausible value) with the parameter estimates and standard errors then pooled according to ‘Rubin’s rules’ (Rubin, 1987). The Stata ‘REPEST’ package, developed by members of the OECD (Avvisati & Keslair, 2014), was used for this analysis.

Economic, social, and cultural status (ESCS) index

The PISA ESCS index is a composite score based on three other indices reflecting parental education, highest parental occupation, and home possessions built via principal component analysis (OECD, 2015). The reliability (Cronbach's Alpha) of the scale was 0.63[11] (OECD, 2017c, p. 340).

Gender

The dummy variable 'female' with assigned values: 0 = male and 1 = female.

Geographic region

This categorical variable describes the location of participant's school, with assigned values: 1 = Greater London, 2 = South, 3 = Midlands, and 4 = North.

School type

This categorical variable describes the type of participant's school, with assigned values: 1 = academy, 2 = maintained selective, 3 = maintained non-selective, and 4 = independent.

Descriptive statistics of all the above variables are presented in Appendix 5.

3.6.9.2 Missing values

To keep the sample size consistent across the models in further statistical modelling, only cases with complete information were used. The small number of students who had missing values in the variables of interest (around 6%) were excluded from analyses, resulting in an analytical sample size of 1,485 cases. More detail on missing values by variables can be found in Appendix 6.

[11] Scale reliability corresponds to sample from the United Kingdom as reported in the OECD technical report (OECD, 2017c).

3.6.9.3 Weights

To account for the complex survey design employed by the PISA study (stratified and clustered sample design), final student weights and balanced-repeated-replication weights are applied throughout the analysis. The balanced-repeated-replication weights are based upon a resampling method and allow the impact of both the stratification and clustering to be incorporated into the estimated standard error[12] (Jerrim et al., 2019). These weights are provided with the data and are recommended by the survey organisers (OECD, 2009). The Stata version 16 (StataCorp, 2019) and the Stata 'REPEST' package (Avvisati & Keslair, 2014) are used to apply the above weights and conduct data analysis.

3.6.9.4 Analyses

Initially, the resulting scores of the CPS competence measures (overall and sub-scales) were compared within student sub-groups based on gender. To explore the degree to which the CPS competence measures are related to similar constructs, correlation analysis was conducted between CPS competence scores and student attitudes towards collaboration measures as well as science, mathematics, and reading performance, in addition to the correlations between CPS competence sub-scales. Following that, regression analyses were conducted to model the relationship between personal (gender and ESCS), contextual (school location and school type), attitudinal (valuing relationships and valuing teamwork), and performance features on students' CPS competence measures (overall and sub-scales). For regression analyses, all continuous (dependent and independent) variables were standardised to have mean zero and a standard deviation of one.

Figures 3.6 and 3.7 present the scatterplots checking for linear relationships between the attitudes towards collaboration scales and the overall CPS competence measure. Preliminary analysis indicated that the relationship between

[12] Alternative methods to account for complex survey designs (e.g., multilevel models) have been argued to ''only capture the impact of clustering and not the impact of the survey stratification per se'' (Jerrim et al., 2019, p. 38).

the overall CPS competence measure and students' attitudes towards collaboration scales is not linear. For that reason, a set of dummy variables referring to quartiles of the attitudes towards collaboration scales are entered into the regression models. Dividing the sample into quartiles facilitates a simple presentation and interpretation of the results, as compared to alternative approaches (e.g., inclusion of a quadratic term), and allows for the detection of more complex patterns in these associations (Jerrim et al., 2019).

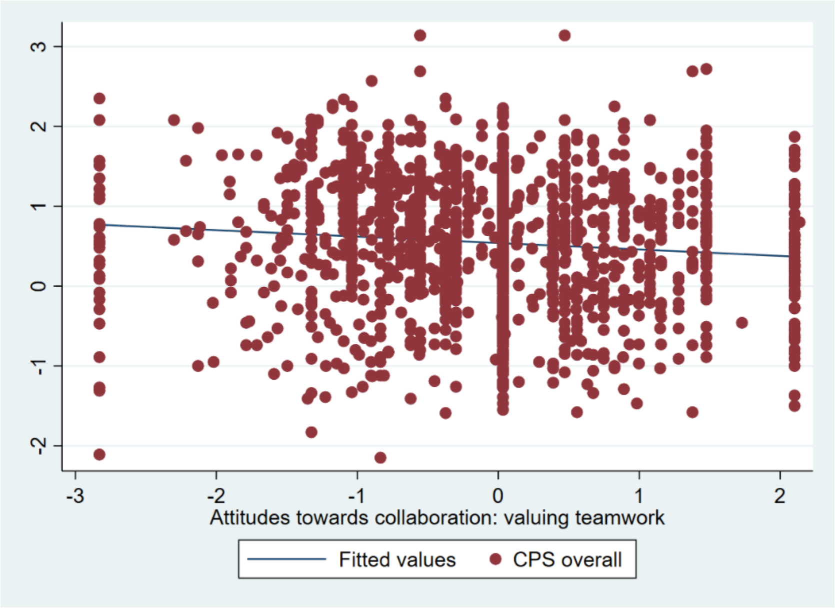


Figure 3.6. Scatterplot of valuing teamwork against overall CPS competence measure

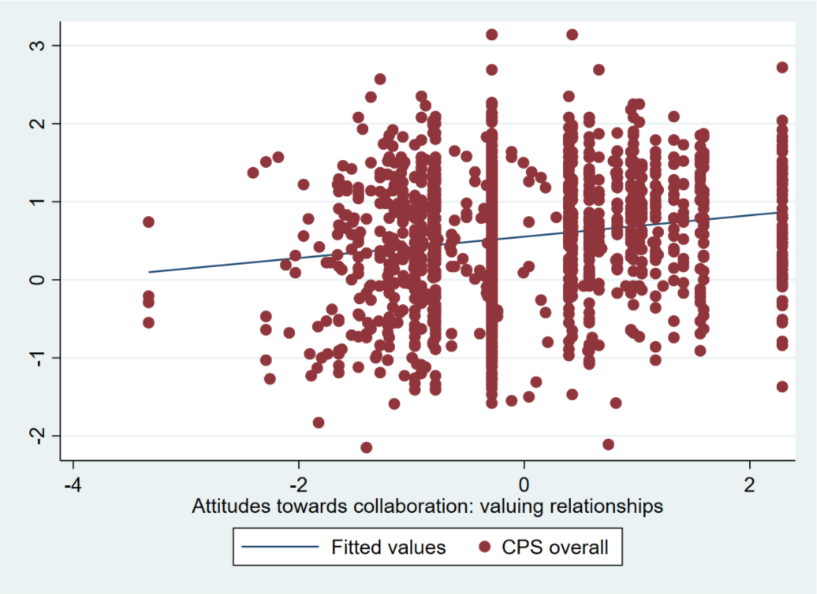


Figure 3.7. Scatterplot of valuing relationships against overall CPS competence measure

Moreover, preliminary analysis indicated potential multicollinearity issues between the mathematics, science, and reading measures (r > 0.70, Appendix 7). To avoid such issues, science performance was selected to be added as a predictor of students' CPS competence[13]. Models were run in the following order to illustrate how parameter estimates changed with the addition of extra variables (Table 3.8). Each model was run four times using each CPS competence measure (one overall and three sub-scales) as outcome variable.

[13] Out of the three PISA subject performances available, science score was added as independent variable in regression analyses since it was the major subject of assessment in the PISA 2015 cycle.

Table 3.8. Outline of models

Explanatory variable	Variable type	Model		
		Model 1a-1d	Model 2a-2d	Model 3a-3d
Gender	nominal	√	√	√
Economic, social, and cultural status	continuous	√	√	√
Geographic region	nominal	√	√	√
School type	nominal	√	√	√
Valuing relationships (quartiles)	ordinal	-	√	√
Valuing teamwork (quartiles)	ordinal	-	√	√
Science performance	continuous	-	-	√

3.7 Methodology for Cognitive interviews with students (Chapter 7)

3.7.1 Overview

Chapter 7 placed the emphasis on students' discourses collected via cognitive interviews. While the PISA 2015 dataset offered information about whether students received credit or no credit (and in few cases partial credit) when responding to the CPS assessment, cognitive interviews provide information about student response processes that were not included in the dataset[14]. As discussed previously (Chapter 2), only a limited range of information was released by OECD regarding what is happening when students responded to the PISA 2015 CPS assessment. Chapter 7 aims to contribute to the validation of the PISA 2015 CPS competence measure using a definition of validity as a unified concept (Messick, 1995). In this sense, it complemented the results of Chapter 6 by investigating what a small sample of secondary school students in England say about how they understood one specific PISA 2015 CPS assessment task[15] and how they explained

[14] Cognitive interviewing was used by the OECD in the instrument development phase of the PISA 2015 study for newly developed material, however, evidence from the interviews conducted is not made publicly available.

[15] OECD released the content from only one CPS task (out of six CPS tasks in total) included in the PISA 2015 CPS assessment. Due to this restriction, this phase of the study makes use of the only publicly available PISA 2015 CPS task named Xandar.

their answers to the CPS items. Therefore, its analysis allowed a more detailed and complex notion of CPS competence to be explored through students' discourses, providing further external validity evidence.

3.7.2 Cognitive interviewing method

Cognitive interviewing (CI) is employed as the main approach to collect qualitative data in Chapter 7. It is a qualitative method that examines the response processes and interpretations of respondents when answering survey questions (Willis, 2005). It is one of the predominant methods for identifying and correcting problems with survey questions/ test items and obtaining evidence on the respondents' question-and-answer process (Beatty, 2004; Willis, 2005). As a method it is used to identify sources of confusion in assessment items and assess validity evidence based on content and response processes (Peterson et al., 2017).

Based mainly on a cognitive four-stage model, proposed by Tourangeau (1984), CI explores the various stages of the question-and-answer process, which are: comprehension, recall, judgment, and response. It has been argued that these are the four major cognitive processes that respondents are presumed to engage in when attempting to answer any item or survey question (Boeije & Willis, 2013; Peterson et al., 2017). Although respondents might not progress through these operations sequentially, it has been suggested that for every item they must understand what the question is asking, retrieve relevant information or knowledge from memory, make a judgment about the item or recalled information, and select a response (Peterson et al., 2017). Additionally, Karabenick et al. (2007) present a conceptual model of six cognitive processes that individuals are assumed to engage in when responding to self-report items:

1. Read and interpret the meaning of words in an item.
2. Interpret what the item is asking and store the interpretation in working memory.
3. Search memory for thoughts, experiences, feelings, attitudes etc. relevant to the content and context of the item.

4. Read and interpret answer response options in the context of the item.
5. Simultaneously think about item, relevant memory, and response options, searching for the answer that most accurately reflects respondent's experience.
6. Select response option that is congruent with information retrieved from memory.

The process described above shows that there are significant demands from the respondents and the actual response selected, in and of itself, has been argued to provide no evidence concerning whether the respondent has actually engaged in the aforementioned steps (Hopfenbeck & Maul, 2011). In addition, it should be noted that, students may not engage in the full reflective response process described due to a simple lack of motivation (lack of personal stakes especially for PISA test) or social desirability. Therefore, CI is used in Chapter 7 to provide evidence of validity based on response processes such as the thought processes involved in responding to an item (American Educational Research Association et al., 2014; Castillo-Díaz & Padilla, 2013; Peterson et al., 2017).

The premise of CI is that intensive interviewing of small numbers of targeted individuals, offers rich insight concerning how survey questions/items provide (or fail to provide) the desired information (Boeije & Willis, 2013). This is achieved due to the way CI is conducted, i.e., asking about the tested survey questions/items, rather than simply collecting answers to those questions/items (Boeije & Willis, 2013). In educational measurement, for example, research studies examining the sources of misfit, and their explanations, have used qualitative case studies to complement statistical methods (Petridou & Williams, 2010a, 2010b). Specifically, Petridou and Williams (2010b) conducted interviews with 'misfitting' pupils (identified by item fit statistics) and their teachers to elicit pupils' and teachers' explanations of statistically unexpectedly correct and incorrect responses in a mathematics test.

3.7.3 Timing for cognitive interviewing

Due to its breadth of application, CI can be conducted at a variety of points in a research study. Although it has been described as a method for item development, CI can also be useful to inform item decisions after (quantitative) data collection (Peterson et al., 2017). Methods such as exploratory factor analysis, confirmatory factor analysis, and Rasch analysis enable researchers to identify poor performing items based on quantitative summaries of data; however, these methods do not explain why an item's statistics are weak (Peterson et al., 2017). For that reason, CI is an appropriate method to use in this phase, following Rasch analysis conducted in Chapter 6.

3.7.4 Cognitive interviewing techniques

The two main techniques commonly used during CI are described as "think-aloud" and verbal probing (Willis, 2005). The "think-aloud" procedure involves asking respondents to describe their thought processes as they answer survey questions/items. When thinking aloud, respondents are asked to report all that comes to mind as they are mentally processing a survey question/item (Boeije & Willis, 2013). The role of the interviewer is to listen and to record but not to comment (Ericsson & Simon, 1993). It has been argued that such verbalisation might feel unfamiliar, and therefore, practice is warranted (Peterson et al., 2017). Furthermore, there is a risk that that even given training in the activity, individuals might simply not be good at thinking aloud when answering survey questions/items (Willis, 2005). The second fundamental technique is for the interviewer to administer verbal probes designed to target the key underlying cognitive processes (Boeije & Willis, 2013). This technique involves the interviewer following up (either immediately or at the end of the interview) student's response to a target item by probing for other specific information (Willis, 2005).

For the purposes of Chapter 7, verbal probing is preferred over "think-aloud" since: i) it tailors the interchange in a way that is controlled mainly by the interviewer, who can focus on particular areas that appear to be relevant as potential sources of response error, ii) it allows examining what may be thought but left unstated, and

iii) it is fairly easy to induce participants to answer probe questions without the need for preliminary training (Willis, 2005). An alternative aim of verbal probing, which is particularly relevant to Chapter 7, is understanding how a question/item works without necessarily seeking to remediate sources of error (Boeije & Willis, 2013). Finally, the age of the respondents was also considered when choosing verbal probing over “think-aloud”. Participants were approximately 15-year-old students, and it was important to make sure that inarticulate or introvert (not outgoing) students would be able to participate in the interviews. “Think-loud” involved the risk for students to find it difficult responding to this open-ended format. Considering the above reasons, in addition to the limited time for each student interview that the school has agreed to give, CI with verbal probing was the best approach to follow.

For CI that relies on verbal probing, a key decision concerns the choice of when to probe and what probes to develop. Verbal probing may be either concurrent or retrospective (Willis, 2005). The difference between these two lies in the sequence with which probes are being asked to the interviewees. In concurrent probing, the interviewer and interviewee take turns asking and answering both target questions and probes. Alternatively, retrospective probing avoids disrupting the interview with probe questions and the interviewer conducts probing at the end, also known as debriefing (Willis, 2005). For this study, retrospective probing was used to create an environment that closely approximated the testing situation and eliminated the disruption in students’ responses (Beatty & Willis, 2007).

Anticipated probes, also referred to as planned, structured, or scripted, are designed prior to the interview to target specific potential areas of confusion (Willis, 2005). It is recommended that the interviewer uses the four cognitive operations, i.e., understanding, retrieval, judgement, and response (Tourangeau, 1984), to appraise each item and anticipate possible misinterpretations (Peterson et al., 2017). Table 3.9 outlines descriptions of the four cognitive operations that were used to appraise items and a list of potential probes adapted from Peterson et al. (2017) and Willis (2005). Spontaneous probes are those probes that emerge as the interview progresses. It might not be possible to anticipate all sources of confusion (Willis, 2005), and the spontaneous probe allows the interviewer to pursue evidence of unanticipated failures in cognitive operations (Peterson et al., 2017). These can be in response to nonverbal indications of confusion (e.g., "You seemed to hesitate there, will you say some more about that?") or can be one of the probes specific to a cognitive operation. For this study, both anticipated and spontaneous probes were used when interviewing students.

Table 3.9 Potential probes

Cognitive operation	Potential probes
Understanding	• What would you say that question was asking of you? • Was there anything confusing about this question? • What do you think it means when it says <*term*>*?*
Retrieval	• How much do you feel you know about <*topic*>? • How easy or difficult is it to remember <*topic*>? • You said <*response option*>. How sure are you of that?
Judgement	• How comfortable did you feel answering this question? • Did it seem like one of the responses is supposed to be the right answer? • Did this question feel awkward or inappropriate?
Response	• Were you able to find your first answer to the question from the response options? • You said <*response option*>. How well did that option accurately reflect the answer you wanted to give? • Was there an answer you wanted to give that was not available? • Were there response options that didn't make sense to you?

Notes: Table adapted from Peterson et al. (2017) and Willis (2005).

3.7.5 The PISA 2015 CPS assessment task ''Xandar''

Chapter 7 makes use of the only publicly available PISA 2015 CPS task named Xandar[16]. In brief, a three-person team consisting of the student test-taker and two computer-simulated partners (Alice and Zach) was asked to take part in a contest where they had to answer questions about the fictional country of Xandar (OECD, 2017b). Information about the 12 items included in the released CPS task Xandar (hereafter described as Xandar items) include:

- the interactive task space that student test takers used,
- content of the pre-defined messages and actions that students could select from,
- the correct action or response to each of the items,
- the skill targeted by each item, and
- justification as to why the action or response is correct.

A detailed description with screenshots for each item can be found in the official PISA 2015 results report (OECD, 2017b) and scoring guide (OECD, n.d.). A summary of the item intent and targeting of each Xandar item is provided in the following section.

3.7.6 Cognitive interview process

CI has been described as a multistep process that involves identification of item intent, data collection, and data analysis (Peterson et al., 2017). These steps will be described in more detail in the following sections. To ensure that crucial pieces of information regarding the methodology employed in this phase are reported appropriately, the guidelines provided in the Cognitive Interviewing Reporting Framework (Boeije & Willis, 2013) are followed.

[16] https://www.oecd.org/pisa/test/other-languages/xandarurlreplacementtest.htm

3.7.6.1 Item intent

One of the first steps involved in CI is the identification of item intent, which directly pertains to the aspect of the construct the item is designed to tap (Peterson et al., 2017). Prior to conducting the interviews, the item intent and the associated CPS skills targeted for assessment by each item were documented for all Xandar items. This description then served as a basis from which to judge if there was a misalignment between how the respondent interpreted an item and what it was intended to measure (Peterson et al., 2017). Table 3.10 presents an overview of the available information about each Xandar item, including the response options for each item and the items' intent. Once item intent is specified, anticipated probes are developed, and the interview protocol is written (detailed in the following section).

For example, in item X11, issues with the response options were anticipated. Specifically, the credited response "We look fine, except for Economy" acknowledges that the team has made progress in some subjects, as shown on the scorecard displayed in the problem space, but that there are still no correct responses in the Economy subject area (OECD, 2017b). However, another response (i.e., "Great, we're half way there") is technically correct, although it does not help the team identify the one area that has not yet been addressed (OECD, n.d.). Probes such as "Were you able to find your first answer to the question from the response options shown?" were developed to determine issues with the response options. Appendix 8 presents the full list of anticipated probes developed for every Xandar item.

Table 3.10. Items included in the PISA 2015 CPS assessment task ''Xandar''

Item	**Item renamed**	**Credited item response option**	**Other item response options**	**Item's intent**	**CPS skill**	**Level of difficulty**
Item 78	X1	Click on the "Join the Chat" button.	Click on other active buttons on the task space.	Item requires student to respond to the directions on the screen.	(C3) Following rules of engagement	Below level 1
Item 79	X2	*3)Maybe we should talk about strategy first.*	*1)I wonder if some of the other teams have started yet.* *2)I hope the questions are easy.* *4)Alice, you can see what to do once we get started.*	Item requires student to take the initiative to suggest the first logical step required to solve the problem.	(C1) Communicating with team members about the actions to be/being performed	Level 2
Item 80	X3	*2)True, but what's a good way to do that?*	*1)Right, the first team to answer all the questions wins.* *3)Do you think all the teams have to answer the same questions?* *4)First we should find out what we'll get for winning the contest.*	Item requires student to focus the discussion on how best to meet the goal of the contest and solicit ideas from the team.	(B1) Building a shared representation and negotiating the meaning of the problem (common ground)	Level 2
Item 81	X4	*4)We can answer more questions if we divide them among us.*	*1)The rules of the contest seem pretty simple. Let's just do our best.* *2)We can each work our fastest, but some of us will still be faster than others.*	Item requires student to volunteer information not specifically requested by the other team members to help the team devise a strategy.	(B1) Building a shared representation and negotiating the meaning of the problem (common ground)	Level 2

			3)It doesn’t matter whether one of us answers more questions than the others, so long as we win.			
Item 82	X5	*1)We could each take one of the subjects.*	*2)If there’s a prize for winning, let’s divide it equally.* *3)The contest lets us come up with our own team strategy.* *4)OK, then we’re ready to begin.*	Item requires student to confirm and slightly extend the approach that has been agreed upon.	(B3) Describe roles and team organisation (communication protocol/rules of engagement)	Level 1
Item 83	X6	*2)Can each of you explain why you want that subject?*	*1)Nobody asked me what subject I want. Why should you guys choose first?* *3)Why are we wasting time arguing about this?* *4)Alice and Zach, are you going to answer questions faster than you choose subjects?*	Item requires student to help team members negotiate a solution when a conflict arises.	(A1) Discovering perspectives and abilities of team members	Level 3
Item 84	X7	*1)It sounds as though People should be Alice’s subject. Zach, are you OK with that?*	*2)Alice, maybe you could study abroad in a visiting students program.* *3)Yes, it’s good to know what your interests are.* *4)People in Xandar probably aren’t very different from people anywhere else.*	Item requires student to evaluate the reasons provided by each team member.	(B3) Describe roles and team organisation (communication protocol/rules of engagement)	Level 1

Item 85	X8	*4)I'll take Geography.*	*1)Well, everyone likes money.* *2)Liking money doesn't mean you understand economy.* *3)We need to stop debating and make a decision.*	Item requires student to assume responsibility for identifying the one remaining subject area that needs to be claimed.	(B3) Describe roles and team organisation (communication protocol/rules of engagement)	Level 2
Item 86	X9	Click on the "Geography" button.	Click on other active buttons on the task space.	Item requires student to act based on the agreed-upon role, respond to directions on the screen, and click the correct button.	(C3) Following rules of engagement	Level 1
Item 87	X10	*4)I should answer the Geography questions. Let's work on the subjects we chose.*	*1)The clock is ticking - let's not waste time on chat messages.* *2)Whoever answered a Geography question, nice work!* *3)Since somebody answered a Geography question, I'm going to switch subjects.*	Item requires student to notice that the event in the problem space violates the agreement that each team member would take one of the subjects.	(D1) Monitoring and repairing the shared understanding	Level 4
Item 88	X11	*3)We look fine, except for Economy.*	*1)I think your scorecard is working – mine is.* *2)Great, we're half way there.* *4)I'm not sure since I don't know the other teams' scores.*	Item requires student to respond to a question from one team member and also provide additional information	(D2) Monitoring the results of actions and evaluating success in solving the problem	Level 4

				about how team is progressing.		
Item 89	X12	*1)Keep trying. When Alice and I are done we'll help you – right Alice?*	*2)Zach, aren't you the one who said we all had to work fast?* *3)Do you expect us to stop what we're doing and help you instead?* *4)Are you behind because you were working on my Geography questions?*	Item requires student to present a proposal that is most effective in working towards the problem solution.	(D3) Monitoring, providing feedback and adapting the team organisation and roles	Level 3

Notes. Table produced using information from OECD (2017b) and OECD (n.d.). Level of difficulty is reported based on the PISA levels of proficiency for the PISA 2015 CPS scale.

3.7.6.2 Data collection

Participant selection

Little research has been conducted on how many interviews are needed when conducting cognitive interviews (Peterson et al., 2017). Recommendations for sample sizes are typically low, ranging from 5 to 15 respondents (Beatty & Willis, 2007; Willis, 2005). Unlike psychometric methods used in establishing evidence of validity, varying perspectives rather than representativeness is the goal when sampling for cognitive interviews (Beatty & Willis, 2007; Willis, 2005). For these reasons, students were selected with the aim to cover a variety of demographic characteristics, social behaviour, and prior attainment.

Cognitive interviews were conducted with 10 students (5 males and 5 females), aged between 15 years-old and 15 years and 7 months old, attending a secondary school in England. The criteria that guided the selection of participants were equal gender distribution, age range close to 15 years-old (similar to the target population of the PISA study[17]), balanced distribution of educational level based on students' grade in the Mathematics and English subjects as described by their teachers, and variety of social behaviour in class as described by their teachers. In addition, it should be clarified that students in England who participated in the PISA 2015 CPS assessment are not the same as the students participating in CI. However, the two student samples share some common characteristics (i.e., age, educational setting, and country) that enable linking the results from the two empirical phases.

Gaining access to potential interviewees and ethics

Mathematics teachers acted as points of contact in the secondary schools. First, an invitation for participation in the research was sent to several secondary schools in Greater Manchester, which had previously hosted students undertaking postgraduate studies (PGCE) at the Manchester Institute of Education (Appendix 9).

[17] PISA assesses students between the ages of 15 years and 3 months and 16 years and 2 months, and who are enrolled in an educational institution at grade 7 or higher (OECD, 2017).

Following that, teachers who responded positively for their school to participate in the study were sent more detailed information including information sheets and consent forms for students and parents (Appendix 10). One teacher responded positively at this stage and selected students (based on the criteria provided) that could be invited for participation. Students' parents/guardians received a letter with information and gave consent to allow their children to be interviewed. All participating students read a letter with participant information and gave assent to be interviewed. The participants were guaranteed confidentiality and that the data would be used solely for purposes related to research. All the participating students are referred to by pseudonyms.

Interviewer's training

Training of new cognitive interviewers has been claimed to be one of the most important, though least documented, aspects of CI (Willis, 2005). For the total number of cognitive interviews conducted in this study, I was the only interviewer. When planning for data collection, I undertook specialised training to develop my skills as cognitive interviewer. In July 2018, I participated in a two-day training in 'Cognitive interviewing for testing survey questions' organised by the National Centre for Research Methods[18]. The training included an introduction to the cognitive stages in answering survey questions as well as the main CI techniques. Most importantly, the training included workshops in which I practised doing cognitive interviews with other participants of the training, using some of the questions included in my interview protocol and received feedback by the trainer.

Conducting cognitive interviews and Interview protocol

Cognitive interviews were conducted on a one-to-one and face-to-face basis in a quiet room that was booked in advance by the mathematics teacher at the participant's school setting. The fieldwork was undertaken between June 2019 and July 2019. The day and time of the interviews was scheduled by the mathematics teacher, and it was within normal school hours.

[18] https://www.ncrm.ac.uk/training/show.php?article=8040

With the permission of student participants and their parents/guardians, interviews were audio- and video-recorded and transcribed for further data analysis. Students completed the Xandar task on a laptop provided by the University of Manchester with their responses being screen recorded. The captured screen-recording video was used for reflection during verbal probing. Each interview was planned to take 40-45 minutes with the time being split between the time students needed to complete the task and the time allowed for verbal probing, considering that longer periods can make excessive demands on attention and motivation of participants (Willis, 2005).

The interview protocol (presented in Appendix 11) consisted of five steps. At the beginning of each interview, effort was made to build a comfortable atmosphere. The interviews were opened by a brief introduction of the purpose of this research. Additionally, I emphasised that there was no right or wrong answer, and that I was interested in the interviewees' own opinions and thoughts. Notes were taken during each interview, which was a useful way to capture the main points and to formulate follow-up questions. In the second step, students were asked to complete the CPS task on the laptop. During that time, their responses were being screen-recorded on the laptop they were using. Once students reached the end of the task I moved to the third step, which involved asking them to explain how they understood every item and how they have gone about answering them. This step was structured to cover two main verbal probes, which were asked to all students for every item:

1) How do you understand this part/statement?
2) Can you explain why you have given this answer?

The first question was used to check item interpretation and the second to check for coherent answer choice (Karabenick et al., 2007). Beyond these core probes, follow-up probes were used when initial responses to the core questions failed to elicit the data needed to effectively assess item performance. A follow-up to the first verbal probe was "Can you tell me a little more about what is happening in this part?” And a follow-up to the second was “Can you tell me a little more about why you chose that answer?” Depending on their response, students were then asked some anticipated probes from the list that has been developed in advance as well as spontaneous probes such as “What do you mean?”. The fourth step included open questions about students’ experience with the assessment in general. Students were free to comment on what they found easy or hard in the assessment or what they would change. Since CPS is not a traditional subject area, in that it is not explicitly taught as a school subject; questions were asked to students with the aim to get an insight into their familiarity with the assessment task. At the close of the interviews (fifth step), students were given the chance to ask questions about their interview or add any last thoughts.

3.7.6.3 Data analysis

There is currently no standard method of analysis for CI (Peterson et al., 2017). Analytic techniques range from less intensive, e.g., making notes as the respondent is speaking, to more detailed coding schemes (Willis, 2005). In Chapter 7, elements of grounded theory and more specifically constant comparative methods are employed. Grounded theory method “uses a systematic set of procedures to develop and inductively derive grounded theory about a phenomenon” (Strauss & Corbin, 1998, p. 24). Grounded theory coding shapes an analytic frame from which the analysis is built (Charmaz, 2006). It consists of at least two main phases: an initial phase involving naming each word, line, or segment of data followed by a focused, selective phase that uses the most significant or frequent initial codes to sort, synthesise, integrate, and organise large amounts of data (Charmaz, 2006). The interview data were analysed through the following stages: familiarisation, reflection, initial coding, focused coding, and axial coding (detailed in the next section). Theoretical sensitivity was a critical part of all these stages. The software NVivo 12 (QSR International Pty Ltd., 2018) was used to support with coding.

Familiarisation

Familiarisation involved reading and rereading data and becoming aware of the main points and details of each individual case. At this stage, interesting response examples or tentative broad codes were briefly noted for each interview. These written records served as reminders of what I have captured during the reading process. Students’ message selections to the CPS items were also recorded as presented in Table 3.11.

Table 3.11. Student responses to Xandar items

	Item (Credited response)											
Interviewee (alphabetical order)	Item X1 (click)	Item X2 (3rd message)	Item X3 (2nd message)	Item X4 (4th message)	Item X5 (1st message)	Item X6 (2nd message)	Item X7 (1st message)	Item X8 (4th message)	Item X9 (click)	Item X10 (4th message)	Item X11 (3rd message)	Item X12 (1st message)
Anna	Click	3	2	4	1	3	1	2	Click	2	3	1
Becky	Click	3	1	4	1	2	1	2	Click	1	3	1
Ella	Click	4	1	1	1	2	1	2	Click	1	1	1
Emily	Click	3	2	4	4	2	1	2	Click	2	2	1
John	Click	3	2	2	1	2	1	4	Click	1	3	1
Leo	Click	2	2	1	1	2	1	2	Click	2	1	1
Maria	Click	4	2	1	1	2	1	4	Click	2	2	1
Oliver	Click	1	1	3	4	1	3	3	Click	3	4	4
Pablo	Click	3	2	4	3	2	1	4	Click	1	2	1
Stephan	Click	3	2	1	1	2	1	4	Click	2	2	1
Total	10/10	6/10	7/10	4/10	7/10	8/10	9/10	4/10	10/10	0/10	4/10	9/10

Notes: Highlighted with grey are the credited responses selected by students.

Reflection

At the stage of reflection, some preliminary cross-case analyses were conducted. This process was carried out by comparing and critically evaluating individual case data with other cases. Corbin and Strauss (1990) state that making comparisons can assist the researcher in guarding against bias and help to achieve greater precision and consistency. During this process, some important questions were asked, such as “Do the ideas in this case differ from other cases?”, and “Are there any new ideas emerging from the case data?”. By asking and answering these questions, similarities and differences among cases were highlighted, and were recorded in memos.

Memo-writing has been argued to constitute a crucial method in grounded theory because it prompts the researcher to analyse the data and codes early in the research process. Therefore, memos were used to catch my thoughts, capture the comparisons and connections, and crystallise questions and directions to be pursued (Charmaz, 2006). It has been suggested that writing successive memos throughout the research process keeps the researcher involved in the analysis and helps to increase the level of abstraction of ideas (Charmaz, 2006). Memos that recorded my early thoughts and reflections with the data, facilitated further systematic coding, and were continuously used throughout the entire process of data analysis. A memo writing example is presented in Figure 3.8. It should be noted that these early stages of familiarisation and reflection happened before the formal coding stages of grounded theory approach described in the following sections.

Memo writing example

In item 87 (X10), John was confused with the response options because he thought he has responded the question about Geography himself. He did not realise the violation in the plan by his teammate and he said that only one response option out of the four made sense to him. *(Being confused -> not realising violation in the plan -> only one response option making sense)*

The item intent was to elicit the skill “Monitoring and repairing shared understanding”, but if the student does not notice the violation in the plan, they might not act towards repairing the agreed plan. Student underperforms when the violation in the agreement is not clear to him. The assumption of this item is that the student will realise that someone else answered his questions. When this does not happen, then three out of four responses do not make sense.

One solution would be for the computer system to show a message making it clear that “Some team member answered the Geography question” or even more explicitly that “Zach answered a Geography question correct”, then the student will have the chance to act towards repairing the shared understanding.

Figure 3.8. Memo writing example

Initial coding

For many grounded theorists, line-by-line coding is the first step in coding (Charmaz, 2006). Line-by-line coding means naming each line of the written data (Glaser, 1978). For this study, more flexibility in the approach was allowed, taking a couple of lines or sentences when necessary to make sense. Initial codes help to separate data into categories, and initial coding frees the researchers from becoming so immersed in the respondents’ views that they accept them without question (Charmaz, 2006). Being critical about the data, as Charmaz (2006) points out, forces the researchers to ask questions about their data.

The following questions were adopted to help me see actions and identify processes in the data (Charmaz, 2006, p. 51):

> -*What process is at issue here? How can I define it?*
> -*How does this process develop?*
> -*How does the research participant(s) act while involved in this process?*
> -*What does the research participant(s) profess to think and feel while involved in this process?*
> -*What might their observed behaviour indicate?*
> -*What are the consequences of the process?*

During the initial coding phase, coding was attempted with words that reflected action. This method of coding restrains the researchers' tendencies to make conceptual leaps and to adopt extant theories before having done the necessary analytic work (Charmaz, 2006). At the same time, a set of concepts from the PISA 2015 CPS theoretical framework, which were identified when reviewing items' intent, were also used as potential codes. Initial codes remained provisional, comparative, and finally grounded in the data (Charmaz, 2006). They were provisional since the aim was to remain open to other analytic possibilities and create codes that best fit the data. In that sense, initial codes were reworded to improve their fit. In short, during this initial coding phase, the aim was to remain open to what the material suggests, stay close to the data, and make codes fit the data rather than forcing the data to fit the codes (Charmaz, 2006).

A total of 80 initial codes were developed during the initial coding phase, with various quotes assigned to each one of them, ranging from one to 40 quotes. The full list of the initial codes is presented in Appendix 12. The following three examples show how interviews were analysed during the initial coding process. Initial codes assigned to quotes are presented in bold and quotes that helped to formulate the codes are underlined.

> Interview excerpt 1: *I would just divide the questions between each other* ***[proposing plan]*** *and then it gets done quicker* ***[getting the task done quicker]****. (Anna)*

> Interview excerpt 2: *I said that we would help him because he was struggling, and when me and Alice have finished, we can help him, so he is not that confused* ***[helping each other]****. Because me and Alice were doing quite well, we only had one question left* ***[evaluating progress]****, so we could help Zach and we could save more time as well, if three of us were doing one subject* ***[saving time]****. (Becky)*

> Interview excerpt 3: *You have to understand, you have to listen to their reasons for why they wanted People [****evaluating reasons****] and allocate which one to each, which one to who you think fits best [****assigning roles****]. From their answers I thought Alice, because she had a genuine reason and a passion for it [****evaluating reasons****].* (John)

Focused coding

Focused coding was the second major phase in coding, after the initial coding, which aimed to help synthesise and explain larger segments of data. Focused coding is more directed, selective, and conceptual than line-by-line coding (Glaser, 1978). It involves using the most significant and/or frequent earlier codes to sift through large amounts of data (Charmaz, 2006). One goal is to determine the adequacy of those codes. It also requires decisions about which initial codes make the most analytic sense to categorise the data (Charmaz, 2006). Initial codes that were only assigned to one reference were likely candidates to be grouped with other relevant codes. A total of 13 focused codes were developed through comparing data to initial codes and across interviews. For example, codes such as 'answering faster' and 'saving time' were grouped together under the focused code 'getting the task done quicker' and reflected the fact that students considered the time limits of the competition to give a response.

Through focused coding, I moved across interviews and compare people's experiences, actions, and interpretations (Charmaz, 2006). Through comparing data to data, I developed the focused code and then, I compare data to these codes, which helps to refine them (Charmaz, 2006). A detailed example of focused coding and the full list of focused codes are presented in Appendix 13 and 14.

Axial coding

Strauss and Corbin (1998) present a third type of coding, axial coding, to relate categories to subcategories. Axial coding is Strauss and Corbin's (1998) strategy for bringing data back together again in a coherent whole. Axial coding follows the development of a major category, although it may be in an early stage of development (Charmaz, 2006). While engaged in axial coding, Strauss and Corbin apply a set of scientific terms to make links between categories visible. They group participants' statements into components of an organising scheme. In one such organising scheme, Strauss and Corbin (1998) include: 1) conditions, the circumstances or situations that form the structure of the studied phenomena; 2) actions/interactions, participants' routine or strategic responses to issues, events, or problems; and 3) consequences, outcomes of actions/interactions. Conditions answer the why, where, how come, and when questions. Actions/interactions answer by whom and how questions. Consequences answer questions of 'what happens' because of these actions/interactions (Charmaz, 2006). The categories that were developed are the following: identifying contextual concerns, showing emotional intelligence, communicating in real life, providing alternative responses, and showing test savviness, which are presented in detail in Chapter 7.

Theoretical sensitivity

Theoretical sensitivity refers to the researcher’s capability to generate concepts from data and develop theory (Glaser, 1992). Theoretical sensitivity may come from the researcher’s professional experience, personal experience, knowledge, and skills (Glaser, 1992; Strauss & Corbin, 1998). To gain theoretical sensitivity, Charmaz (2006) points out that researchers look at studied life from multiple vantage points, make comparisons, follow leads, and build on ideas. The acts involved in theorising foster seeing possibilities, establishing connections, and asking questions. Consistent with Glaser's (1978) guidelines, Charmaz (2006) stresses using gerunds (i.e., a verb form which functions as a noun, ending in '-ing') in coding and memo-writing. Gerunds prompt thinking about actions, and therefore, it is suggested to emphasise on actions and processes, not on individuals, as a strategy in constructing theory and moving beyond categorising types of individuals (Charmaz, 2006).

To conduct data analysis, I developed my theoretical sensitivity as a researcher in several ways. Firstly, I built a strong theoretical framework by reviewing how CPS has been conceptualised in the literature and by reviewing extended literature in the field of CPS more generally (Chapters 4 and 5). This helped me to think about the interview data in theoretical terms. Secondly, following recommendations by Charmaz (2006), gerunds were used when coding the interviews, focusing on the actions of the respondents. Finally, discussions took place with the supervisory team about problems encountered in the coding process as well as feedback from various analysis stages.

3.8 Reliability and trustworthiness

As argued by Sammons and Davis (2017), providing a clear and sufficiently detailed description of the mixed methods design is vital in judging the rigour of a study and the robustness of the knowledge claims that it makes. This chapter has clearly documented and explicitly explained the research design and various methods used, providing justification for the decisions made in every aspect of this study. Following from that, when frameworks and relevant guidelines were used to guide the validation process including the data collection and analysis procedures, these were described and referenced appropriately to enhance the transparency and replicability of the study. In the following chapter (Chapter 4) the first research paper of the thesis on the conceptualisation of CPS, following a systematic literature review methodology, is presented.

Chapter 4 Conceptualisation of collaborative problem solving: a systematic literature review

4.1 Abstract

The article presents the results of a systematic review of the literature on collaborative problem solving (CPS) in education published in research journals over the past two decades. In the research field of education, arguments have been made that the conceptualisation of CPS was inconsistent making the literature as a whole incoherent. This article summarises how the concept was employed by analysing the definitions, theoretical underpinnings, research purposes, and methods for data collection and analysis adopted in existing literature. A total of 59 articles from 34 different journals were deemed relevant for review based on a set of inclusion criteria introduced. The analysis revealed three main categories of focus: i) CPS competence in which CPS was defined as a collaborative competence utilised in a problem-solving process, ii) CPS practice in which CPS was seen as a type of pedagogic approach/intervention providing the social context in which learning took place, and iii) CPS interaction in which CPS was conceived as an activity taking place in a joint problem space. The empirical work was found to be using two main units of analysis, focusing on two distinct levels of description: (i) a single utterance used to reflect individual cognition which focussed the analysis on the individual level and (ii) a sequence of multiple utterances by different individuals used to examine social interaction within the group focusing the analysis at the group level. In conclusion, it is suggested that the newly developed assessments of CPS competence will need to be examined and operationalised considering issues of validity and authenticity. Finally, future research will need to focus on interaction processes from different levels of description reflecting the collective, extending existing conceptual frames and in relation to the purpose of research and its epistemology.

Keywords: collaborative problem solving, systematic literature review, theory, definition, conceptualisation

4.2 Introduction

Numerous studies in educational research have recently used the concept collaborative problem solving (CPS) to understand some aspects of learning and performance of problem solving, a focus of much debate in education for about a century (e.g., Dewey, 1933; Mayer, 1992; Newell & Simon, 1972; Pólya, 1945). Its recently increased popularity is evidenced by the growing number of policy documents, disseminated by organisations and initiatives such as the Organisation for Economic Co-operation and Development (OECD, 2017a) and Assessment and Teaching of 21st Century Skills project (Binkley et al., 2012), identifying CPS as a set of skills needed to navigate our global society. In addition, assessments of CPS skills have been recently developed, e.g., the 2015 Programme for International Student Assessment (PISA) (OECD, 2017b), in an attempt to measure, and consequently evaluate, student skills to meet the CPS demands of their future careers. Acknowledgement of the importance of CPS has also motivated interest in curriculum reform in different parts of the world, where education systems have progressively begun to incorporate CPS into their curricula and teaching approaches (Care, Anderson, et al., 2016).

Given its increased popularity and perceived importance, researchers have highlighted the need for consistent and coherent definitions as well as more refined methodological approaches (e.g., Andrews-Todd & Kerr, 2019; Sun et al., 2020; von Davier & Halpin, 2013). However, despite the repeated calls for conceptual coherence, there is no widely accepted definition of CPS in the literature. This has some obvious challenges and implications for the development of the field including, for example, the difficulty of evaluating a poorly defined concept and the risk of oversimplifying a complex concept for assessment purposes. A few studies have so far reviewed the literature related to the CPS concept for distinct purposes (Cukurova, Luckin, & Baines, 2018; Graesser et al., 2018; Oliveri et al., 2017). These reviews, although valuable, are limited in at least three ways: First, they have been informed by certain conceptualisations of CPS taken by the researchers, and thus, they ended up reviewing different bodies of literature which use inconsistent conceptions. Second, they solely focused on certain educational settings (e.g., higher education or secondary education only); another limitation that this paper will overcome by extending the context of reviewed studies. A third limitation is the lack of acknowledging the diverse meanings of CPS concept that this paper will uncover. There is, therefore, a need for a systematic study on how the concept of CPS has been defined and used in the problem-solving literature to date. To fill that gap, the aim of this article is to systematically examine and analyse how CPS, as a "concept" in research practice, has been conceptualised and operationalised in recent empirical and theoretical educational research, for the purposes of informing (i) future research (including meta-analyses and reviews), and (ii) policy and practice about the current state-of-the-art in the field.

The next section presents the background including its research questions, followed by an overview of the systematic review methodology adopted here. The article continues with the results, followed by a discussion and conclusion.

4.3 Background

4.3.1 Historical perspectives and theoretical grounding

As a term, CPS brings together "collaboration" (including collaborative learning), and "problem solving", both of which are complex concepts on their own. These concepts have a substantial research history, each worth overviewing, to ground CPS within the wider literature. Problem solving refers to cognitive processing directed at achieving a goal without knowing a solution method (Mayer, 1992, 2013; Mayer & Wittrock, 2006). Such a definition is broad enough to include a wide array of cognitive activities. Collaboration means "work" ("labour") together with others ("co") whether in play, work, or education. Collaborative learning, which is broadly defined as a situation in which two or more people learn or attempt to learn something together (Dillenbourg, 1999), has become an increasingly important part of education (O'Donnell & Hmelo-Silver, 2013). The next section overviews the main theoretical approaches that shaped the study of problem solving, and collaborative learning, respectively.

4.3.1.1 Classic theoretical approaches to the study of problem solving

Interest in the study of problem solving is not new. In the classic book *How We Think* (1910, 1933) John Dewey explored the process of reflection in relation to the scientific inquiry into a problem. The heart of Dewey's analysis was the description of the reflective thought model, in which, reflection on solving a problem involved a sequential process with a specific target in mind (Farra, 1988). The stages and processes that are the focus of research on problem solving today, e.g., categorisation, coding, decision making, and judgement, were also recognised in Dewey's phases of reflective thinking (Dominowski & Bourne, 1994).

Major theoretical approaches developed throughout the 20th century include associationism, Gestalt psychology, and information processing. According to associationism, cognitive representations in the mind consist of ideas and links between them and cognitive processing in the mind involves following a chain of associations from one idea to the next (Mayer, 1992). The main representative of this approach, Thorndike (1911), viewed solving a problem as simply a matter of trial and error and accidental success. However, a major challenge in associationism concerns the nature of transfer, i.e., explaining where a problem solver finds a creative and novel solution not performed before (Mayer, 2013). The Gestalt approach to problem solving was developed in the 1930s and 1940s as a counterbalance to associationism (Mayer, 2013). Learning was regarded as a process of recognising relationships and developing insights (Schoenfeld, 1985). According to this approach, cognitive representations consist of coherent structures and the cognitive process of problem solving involves building a coherent structure (Mayer, 2013). The reliance of this approach on the subconscious led to it being challenged due to the lack of reliability and validity in methodological implementation (Schoenfeld, 1985).

In 1945, Georg Polya's book *How to Solve It* marked a turning point in the study of problem solving and the implementation of problem-solving approaches to mathematics education (Schoenfeld, 1987). His famous four-phase description of the problem-solving process included understanding the problem, devising a plan, carrying out the plan, and looking back. Years later, in the 1960s and 1970s, the information processing approach to problem solving was developed, based on the influence of the computer metaphor, i.e., the idea that humans are processors of information (Mayer, 2009). In their book *Human Problem Solving*, Newell and Simon (1972) tested their conceptions of human problem solving using a computer simulation as a research method as a demonstration of the wider applicability of certain general problem-solving techniques. The information processing approach is limited in usefully describing problem solving for well-defined rather than ill-defined problems (Mayer, 2013).

Some of the aforementioned approaches were also influential in the most recent study of problem solving. For example, drawing on the information processing approach and Polya's ground-breaking work, problem solving was defined in the PISA 2012 study as: ''an individual's capacity to engage in cognitive processing to understand and resolve problem situations where a method of solution is not immediately obvious. It includes the willingness to engage with such situations in order to achieve one's potential as a constructive and reflective citizen'' (OECD, 2013, p. 122). This definition was developed for the purposes of assessing students' individual problem-solving competence in the context of international comparisons and has been used later to develop the PISA 2015 CPS framework.

4.3.1.2 Classic theoretical approaches to the study of collaborative learning

Much of the research on collaborative learning is rooted in the work of Piaget and Vygotsky (Dillenbourg et al., 1996). Specifically, Vygotsky's (1978) concept of the zone of proximal development is often used to explain that collaborative learning has benefits for learners since the more capable learner (including the teacher) can help the less capable learner to accomplish a task (Janssen et al., 2010). Many also see the more capable being enabled by the less capable, by being enabled to engage in explanations that require them to reflect on their processes. Vygotsky also saw the emotions as intimately entwined in collaborations, the "alpha and omega" of learning.

The long and rich tradition of collaborative learning has led to a vast number of research studies, including reviews and meta-analyses, examining the effects of collaboration on learning outcomes of individual students, such as student achievement (e.g., Roseth et al., 2008; Slavin, 1980). This line of research has become known as effect-oriented research and has been criticised for overly focussing on learning outcomes and not studying the interaction process itself, nor the intervening variables that may affect the outcome of collaborative learning, making it difficult to explain the variability in research findings (Cohen, 1994; Dillenbourg, 1999; Dillenbourg et al., 1996).

A second research approach, known as process-oriented research, focused on understanding the underlying mechanisms of collaborative learning such as the complex relationships and interactions between the task, the learner, and the group characteristics (Dillenbourg et al., 1996; Janssen et al., 2010). However, to date, only a limited number of studies have taken a process-oriented view, and thus, little is known about the dynamics of collaborative activity (Seidouvy & Schindler, 2019). Finally, a third strand of research on collaborative learning has emerged, commonly called computer-supported collaborative learning, which is concerned with studying how people can learn together with the help of computers (Stahl, 2013). Such an approach focuses on new technologies for mediating, observing, and recording interactions during collaboration (Lai, 2011). With the advent of computer-supported collaborative learning, interest changed from assessing individual student outcomes to analysing the group processes, and therefore, researchers have begun to consider group-level conceptualisations, such as group cognition (e.g., Stahl, 2006).

There were prior attempts to understand these processes, especially in mathematics education: Cobb and Gravemeijer (2008) for instance focussed attention on the growing collective practices/knowledge at the level of the classroom, including assumptions about what is ‘taken as shared’ e.g., what needs to be shared, but also what can be taken for granted in the class, and so did not (normatively) need to be said.

Recognising human cognition to be a social product of interaction among people, post-cognitive theories, with a focus on artifacts, communities of practice, group cognition, activity, and mediations by actor-networks have also been established and studied (Stahl, 2013). Instead of viewing knowledge as mental representation of individuals, these theories consider small-group processes (Stahl, 2006), embodied habits (Bourdieu, 1977), activity structures (Engeström, 1999), and community practices (Lave, 1991). To sum up, collaborative learning can be conceptualised in different ways that have implications for investigating collaborative activities.

4.3.2 Current state of the literature

To identify recent research reviews related to the definition of CPS concept, I searched the Scopus database (in title, abstract and keywords) and the University of Manchester Library database (in title) using the terms "collaborative problem solving" and review/concept/theory[19]. Through this search, I found only three major research reviews of CPS in the field (Cukurova, Luckin, & Baines, 2018; Graesser et al., 2018; Oliveri et al., 2017). A closer look at them revealed that what was presented as evidence from the field was aligned with a specific view of CPS taken by the researchers. More specifically, their reviews were informed by a distinct aim, such as the assessment of a learning outcome (Oliveri et al., 2017) or an evaluation of a pedagogical practice (Cukurova, Luckin, & Baines, 2018), resulting in reviews examining different bodies of the literature.

In the (chronologically) first review, Oliveri et al. (2017) aimed to inform the assessment of individuals' CPS competence, and they defined CPS as a collection of skills that are arguably important for daily life, work, and schooling in the 21st century. They organised these skills in four components: teamwork, communication, leadership, and problem solving. This review limited the discussion of studies to those from higher education and workforce contexts only. Graesser et al. (2018) defined CPS as an essential skill in the home, the workforce, and the

[19] The search was limited to studies published in English since 2017.

community, that requires both cognitive and social skills. Their review described how the concept was defined and operationalised in two frameworks (Hesse et al., 2015; OECD, 2017a), both defining CPS as an individualised competence of secondary school students. Finally, Cukurova et al. (2018) investigated the effectiveness of CPS as group work pedagogy by examining existing reviews and meta-analyses. Interestingly, no reviews or meta-analyses specifically targeting CPS were found, and therefore, the review was limited to discussing studies embedded within the broader sphere of collaborative and/or cooperative learning.

There are several challenges associated with the lack of a conceptual coherence in CPS research with implications for the development of the field. I argue that if the concept is not clearly defined, this could critically threaten how the richness of such a complex "real life" phenomenon is captured and raise concerns about authenticity and external validity. It is not obvious, for example, that CPS in the real world of work or the home is well represented by its reduction to a list of skills to be tested in simulated conditions in schools, and even less in individual performance in a computer simulation. Finally, without a common language to describe CPS, it can be challenging for researchers, practitioners, and policy makers to communicate and debate about it.

4.3.3 Research questions

Existing reviews do not yet provide an in-depth insight into the variety of conceptualisations of CPS in the education field and across educational settings. Thus, this article aims to overview CPS conceptualisations within educational research to provide a holistic picture of the diverse use of the concept, for the purposes of informing future research (including meta-analyses and reviews) and, policy and practice, about the current state-of-the-art in the field. To accomplish such a task, a systematic literature review methodology is used, where CPS definitions, theoretical underpinnings, research purposes, and methods for data collection and analysis, are gathered and analysed.

The following research question and sub-question guide the review:

> RQ1: How has “collaborative problem solving” been conceptualised and operationalised in the educational research community?
>
> > RQ1.a: How are the variations in the CPS conceptualisations explained by diverse research purposes?

4.4 Methods

4.4.1 Systematic review methodology

A review of research is a form of research in itself, and a systematic review, like the one herein, is ‘systematic’ in the same way that any empirical research needs to be systematic and transparent, so that the results can be interpreted and assessed in the light of how they were produced. A pre-defined procedure proposed by Gough et al. (2016) and Petticrew and Roberts (2006) to guide systematic reviews including the following steps was employed: developing research questions, determining the types of publications to be located, carrying out a comprehensive literature search, formulating inclusion criteria, appraising study quality, extracting data and synthesising.

4.4.2 Literature database: article search approach and inclusion criteria

To gather relevant evidence, an electronic search was conducted in four scientific databases: SCOPUS, Web of Science, Education Resources Information Centre (ERIC), and British Education Index (BEI). These databases were chosen as they offer an extensive coverage of research literature in the social sciences and two of them (i.e., ERIC, BEI) are relevant to educational research. This review was focussed on the field of education, since students emerging from schools into the workforce and public life are expected to be able to work in teams to solve diverse problems (Rosen & Foltz, 2014). The search was first conducted in January 2019 and was updated in April 2020 to include the most recent publications. The terms “collaborative problem solving” and “student(s)”, “pupil(s)” or “learner(s)” were used in the search.

The search targeted research literature which focused specifically on the CPS construct rather than related terms, such as “teamwork”, and “problem solving”. Boolean operators were employed to combine the key terms as follows: “collaborative problem solving” AND (student* OR pupil* OR learner*), making the search specific to student populations. The search terms were applied to the fields of title, abstract, and keywords.

Articles were included based on the following criteria:

- peer-reviewed journal articles,
- published between 2000 and 2020,
- written in the English language,
- full text available,
- referred to student populations from various educational settings (from reception to higher education),
- were concerned about CPS in the field of education,
- provided a clear definition/conceptualisation/framework of CPS.

The publication period was restricted to the last two decades aiming to map current literature of the 21st century. Search was limited to articles published in peer-reviewed journals to assess study quality, although it is recognised that this has its own limitations as a quality check criterion (Alexander, 2020). A publication was retained for review when CPS-related research formed part of the content and focus of the article, meaning that CPS was not used merely as an example among other aspects of learning, and CPS was not used as a term specific to a field other than education. In this way, only articles with a focus on CPS, targeting student populations and providing explicit evidence for their conceptualisation were considered as relevant for inclusion in the review.

4.4.3 Selection of articles

To screen and select articles to be included in the literature review, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) method (Moher et al., 2009) was used (see Figure 4.1). Publication records were managed in the EPPI-Reviewer 4 (Thomas et al., 2010) software. The article search yielded 702 articles in total. Removal of duplicates resulted in 374 unique articles to be screened. Screening was conducted in two rounds: 1) applying the inclusion criteria to titles and abstracts, and 2) applying the inclusion criteria to the full text of the remaining articles (Figure 4.1). If a decision for inclusion could not be made by reading only the title and abstract, then the article was included in the sample for further screening on the full text.

To ensure that important and relevant research evidence has not been missed, a targeted search was conducted. This involved examining the reference lists of existing literature reviews (i.e., Cukurova, Luckin, & Baines, 2018; Graesser et al., 2018; Oliveri et al., 2017) and the publication records of authors, who were found to frequently contribute to the topic. This resulted in identification of five additional articles, bringing the total to 59 articles for inclusion in this review (see Figure 4.1; the full reference list is included in References). Randomly chosen articles (n = 40) were reviewed by the second author and there was a 95% match in the ratings (i.e., inclusion and exclusion) of articles by both reviewers. Any disagreements were resolved through discussion.

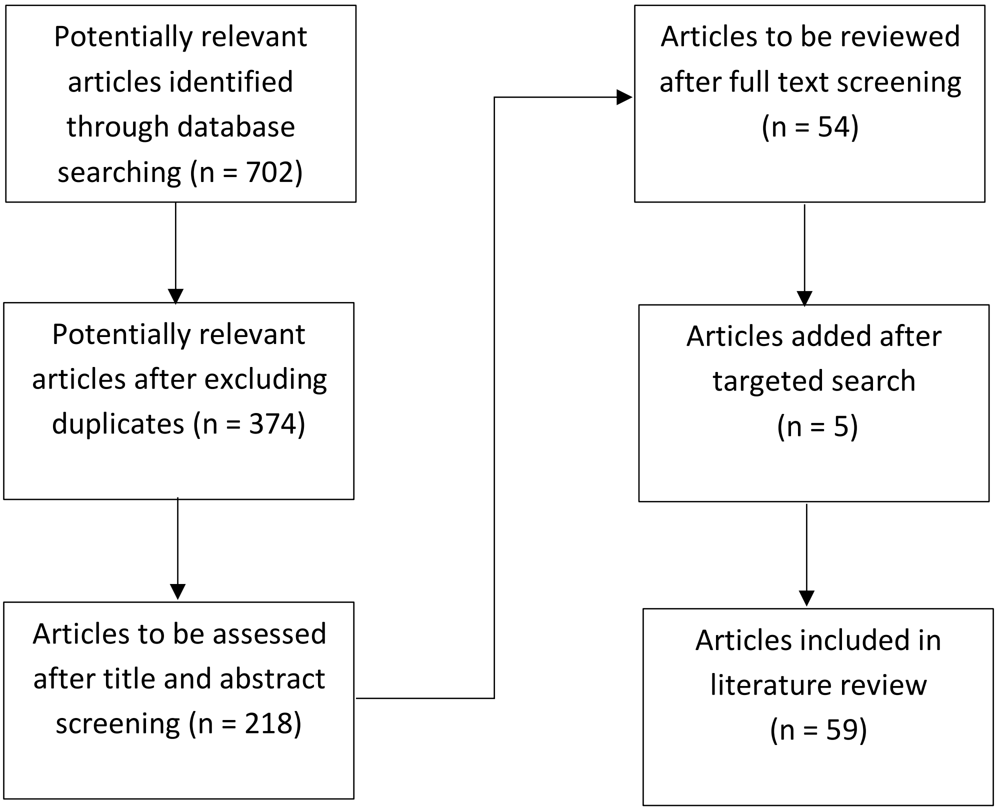

Potentially relevant articles identified through database searching (n = 702)
Potentially relevant articles after excluding duplicates (n = 374)
Articles to be assessed after title and abstract screening (n = 218)
Articles to be reviewed after full text screening (n = 54)
Articles added after targeted search (n = 5)
Articles included in literature review (n = 59)

Figure 4.1. Flow diagram of systematic literature review selection

4.4.4 Analysis

To make sense of the literature, information is summarised in a form that can be easily viewed, analysed, and managed, in other words, the literature is coded. Applying descriptive codes to articles enables the description or mapping of the size and nature of the literature before examining it in depth (Gough et al., 2016). Throughout the analysis, the unit of analysis was the article and coding was based on the following descriptive variables:

- General information: authors, year of publication, journal name, country affiliation of the institution of the first author, type of research (empirical or theoretical).

- Research design (for empirical research only): evidence type, research aims, educational setting, sample size, curriculum area, group composition.

Thematic synthesis was employed (Thomas & Harden, 2008) to analyse definitions of CPS (and the associated operationalisations) from each article. As a first step, each article was read to gain insight into how CPS was understood and used by the author(s), looking for clear statements about CPS definitions. These were extracted in the form of literal quotes, along with theories consulted or used, research purposes and methodology. The extracted raw data was coded focusing on identifying defining features of CPS conceptualisation that appeared to be important for the authors. This first step remained closely attached to the data itself, using constant comparison to check consistency of interpretation (Thomas & Harden, 2008).

In step two, descriptive themes were constantly refined to reflect the various defining features of CPS conceptualisations. In the final step, I have gone beyond the content of the original articles by using the descriptive themes to answer my research questions and build more abstract categorisations (Thomas & Harden, 2008). The main result of this final step was the categorisation of articles into three different groups based on the distinct foci guiding their conceptualisation. A description of their purpose and operationalisation was also produced considering the information extracted from each article's stated research purpose and methodology. Final categories and their descriptions are presented in the results section below.

4.5 Results

This section presents the general characteristics of the reviewed articles, followed by the conceptualisations identified. The final dataset includes 59 articles that matched the inclusion criteria.

4.5.1 Descriptive statistics of articles

Research published on CPS shows an increasing trend over the last two decades, and especially over the last 10 years, confirming that CPS is very topical within educational research (see Figure 4.2). The rise of publications observed in 2017 could be attributed to the release of PISA 2015 results (November 2017) on the assessment of students' CPS competence, which is likely to have increased attention towards CPS. It is also likely that the drop in 2020 is artificial, as publications are not representative of the whole year.

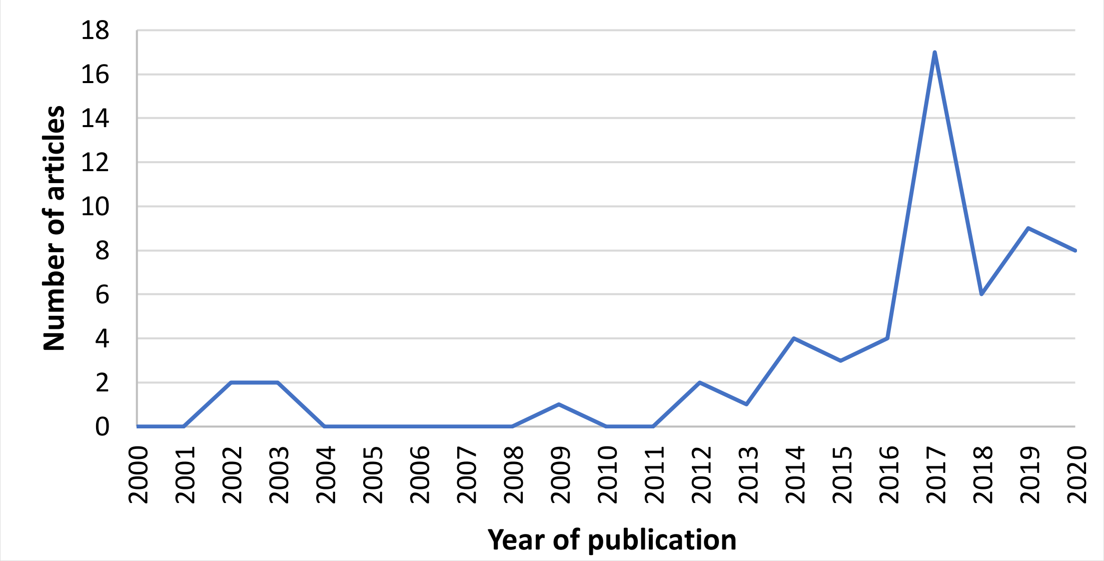


Figure 4.2. Publications per year

Most articles on CPS have been published by authors affiliated with institutions in the following countries: USA (n = 21), Australia (n = 10), and Taiwan (n = 8). The high number of publications in those countries is linked to specific initiatives and policy decisions. Specifically, in 2008, three large technology corporations (Cisco, Intel, and Microsoft), together with six governments (including Australia and USA), funded the project Assessment and Teaching of 21st Century Skills (Griffin & Care, 2014). In 2014, after the announcement of the PISA 2015 focus on students' CPS performance, Taiwan's Ministry of Education and Ministry of Science and Technology, launched several projects that would allow for a nationwide investigation of CPS skills (e.g., K.-Y. Lin et al., 2015). Authors of the remaining articles are affiliated with institutions in 11 different countries (see Appendix 15 for coding information from the full list of articles).

The articles were published in 34 different journals. The journal containing the most articles is *Computers in Human Behavior* (n = 8), followed by *Computers and Education* (n = 5). It is interesting to note that more than half of the articles (n = 34) were featured in journals related to Information and Communication Technologies, which might be related to the increasing use of technological tools for the facilitation of teaching and assessment of CPS.

Thirteen of the articles reviewed are theoretical pieces of work (including literature reviews). Of the empirical articles (n = 46), most employed quantitative methods (n = 43), while very few employed qualitative (n = 2) and mixed methods (n = 1). Sample sizes ranged between 6 and 52,110 students. Most of the empirical articles explored CPS in secondary school settings (n = 35). The remaining explored CPS in primary school (n = 5), higher education (n = 5) or a mix of educational settings (n = 1). No articles were concentrated on students at early years educational settings. The subject or curriculum area specified in most of the empirical articles was Science, Technology, Engineering, and Mathematics (STEM, n = 22), followed by Economics (n = 2), Geography (n = 1), and Language (n = 1). In addition, 17 articles used content-independent tasks that did not require specific subject knowledge from students. Four articles included both subject-specific and content-independent tasks. Most of the empirical articles placed students in groups composed of human participants only (n = 34), while 8 articles placed students in groups with computer-simulated partners, mostly employing the PISA 2015 CPS assessment approach and framework. Finally, four articles employed both group formats.

4.5.2 Conceptualisations of collaborative problem solving

Based on the definitions, research purposes, and methods reported by the authors, the literature on CPS can be broadly divided into three categories of focus: i) CPS competence, ii) CPS practice, and iii) CPS interaction. The distribution of the articles based on these categories of conceptualisations is presented in Table 4.1. In CPS competence, researchers aimed to assess students' CPS skills, based on observed behaviours, using individualised assessments. In CPS practice, researchers aimed to evaluate the effect of CPS as a type of pedagogic approach/intervention on student learning outcomes (whether this was their CPS skills or other knowledge and skills) or attitudes. In CPS interaction, an emphasis was placed on how students interacted with each other during CPS activity.

Table 4.1. Distribution of articles according to their conceptualisation of CPS

Category	Conceptualisation	Operationalisation and preferred methodology
CPS competence (n = 36, 61%)	CPS as an individual collaborative competence utilised in a problem-solving process	-Competence inferred from students' observed behaviours during the problem-solving process -Data collection mainly through computer-based tasks -Quantitative methods of data analysis -Empirical articles n = 26, theoretical articles n = 10
CPS practice (n = 17, 29%)	CPS as intervention affecting student outcomes and attitudes	-Educational intervention evaluated based on student outcomes -Data collection mainly through pre-, and post-tests -Quantitative methods of data analysis -Empirical articles n = 14, theoretical articles n = 3
CPS interaction (n = 6, 10%)	CPS as inter-activity	-Patterns of interaction during the problem-solving process -Data collection through computer-based or face-to-face problem-solving tasks -Qualitative and quantitative methods of data analysis -Empirical articles n = 6

4.5.2.1 CPS competence

The first group of articles (n = 36, 61%) was characterised by a conceptualisation of CPS as an individual's collaborative competence utilised in a problem-solving process, aimed to be mainly individually assessed, even if in a group problem-solving context. Amongst these articles, CPS was generally defined as a "set of skills", including both social and cognitive skills. In general, authors emphasised individual performance in a collaborative event, as opposed to group performance, and made inferences about individual students' ability based on their observed behaviours (Scoular et al., 2017). Overall, the focus of articles in this group was centred on test development and scale validation.

A widely cited definition of CPS was provided by the PISA 2015 CPS framework (OECD, 2017a). Building on a previously developed definition of individual problem-solving competency in PISA 2012 and existing literature on problem solving (e.g., Griffin et al., 2012; Mayer, 1992; Newell & Simon, 1972; O'Neil et al., 2003; Pólya, 1945), PISA 2015 defined CPS as: "the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution" (OECD, 2017a, p. 134). An 'agent' could be either a human team member or a computerised simulation of a human team member. For the PISA 2015 CPS assessment, the latter was used in an attempt to control the collaborative interaction. This definition was developed for a specific purpose, i.e., to create a standardised large-scale assessment for 15-year-old students' CPS skills and subsequently provide cross-country measures of student performance. According to the definition, CPS competence involved an individual's cognitive processing that engaged both cognitive and social skills needed to solve the problem that did not rely on specialised content knowledge (OECD, 2017a). In this way, the idea that CPS skills were a 'generic' type of skills that could be applied across disciplines, or in fact in discipline-free problems, was endorsed. Cognitive processes could then be inferred from the actions performed by the individual, communications made to others, and final products of the problem-solving tasks.

Another widely cited framework of CPS was the Framework for teachable CPS skills, developed by Hesse et al. (2015) for the purposes of the Assessment and Teaching of 21st Century Skills project (Griffin et al., 2012). It defined CPS as "a set of skills on which individuals need to rely when the capacities or resources of just one person are not sufficient to solve a problem" (Care & Griffin, 2014, p. 371). Similar to the PISA framework, the focus was on the skill of the individual in the collaborative partnership, as opposed to a focus on the collaborating pair (Care, Scoular, et al., 2016). These frameworks also formed the theoretical foundation for the development of other CPS frameworks, e.g., Higher education readiness CPS taxonomy (Oliveri et al., 2017), CPS ontology (Andrews-Todd & Kerr, 2019), CPS Generalised competency model (Sun et al., 2020). A common feature across these frameworks is the analysis of CPS into contributing skills, frequently organised in (at least) two dimensions – a social and a cognitive dimension. Social skills helped students coordinate actions in synchrony with other participants, while cognitive skills referred to the ways in which problem solvers managed the task at hand and the reasoning skills employed (Hesse et al., 2015). Such a step in the description of CPS competence construct was argued to be necessary for subsequently facilitating the design of assessment tasks and measurement of CPS competence (Care, Scoular, et al., 2016).

Another common feature in the articles of this group is the condition of interdependency describing the problem-solving tasks utilised as assessment tools. This condition develops situations where collaboration is necessary to performing successfully on a task (Rosen, 2015). One way to achieve this is by allocating asymmetrical roles to students, where no one has all the information or resources required to solve the problem. These tasks, also known as hidden-profile or jigsaw-type problems, were argued to restrict individual activity, and through the dissemination of resources and information, they prompt collaboration (Scoular & Care, 2019). For example, in a computer-simulated assessment task, students were asked to secure a warehouse by correctly positioning the minimum number of security cameras required and they needed to work together to find out how the cameras operated (Harding et al., 2017). One student had control of the cameras but could not see the areas the cameras covered, while the second student could see the areas covered by the cameras but not the cameras themselves.

Interestingly, almost all empirical articles (25 out of 26) utilised computer-simulated, scenario-based tasks for the assessment of students' CPS skills. As students were working with others to solve a problem, their actions and communications, and their sequences, were captured throughout the problem-solving process in time-sequenced log files. Authors then made inferences about individual students' skills, based on their observed behaviours, as represented through the log files generated (e.g., Scoular & Care, 2020). Individual cognition within the group was prioritised by taking a single utterance as the unit of analysis and for scoring. This was defined as any observable behaviour (e.g., chat message, clicking a button) made by the individual to change the state of the problem. Each single utterance was coded as indicative of a CPS skill.

A considerable number of empirical articles (9 out of 26) utilised computer-simulated partners to provide more control over the collaborative interaction. As a result, the communication between team members was constrained and students were forced to choose from a short list of predefined chat messages to communicate (e.g., Herborn et al., 2017; K.-Y. Lin et al., 2015; Polyak et al., 2017; Rosen, 2015). As expected, concerns were raised about the validity and authenticity of such highly constrained approaches (Graesser et al., 2018; Scoular et al., 2017; Webb & Gibson, 2015). Only one article was found to facilitate face-to-face collaboration among students, although it did not analyse students' verbal interactions (Cukurova, Luckin, Millán, et al., 2018). Instead, measures of non-verbal behaviour (e.g., hand position, head direction) were examined.

Generally, the focus of this group of articles was on the individual level, however, a few articles proposed new methods, such as a pathways approach (Vista et al., 2017) or defining individual and group indicators (Cukurova, Luckin, Millán, et al., 2018; Yuan et al., 2019) to assist with the investigation of the CPS competence at both the individual and group levels. Finally, quantitative content analysis was commonly used (9 out of 26 articles) as a method of data analysis, counting the codes or observer ratings attributed to individuals' actions and communications during the problem-solving process. The attempt to enforce a category to each fixed unit without considering how students sequentially organise their actions in the environment, has been argued to be too restrictive to adequately capture the complexity of the interaction (Stahl, 2006). The second most commonly used method of data analysis (7 out of 26 articles) was item response theory for the purposes of measure construction and validation.

Summarising, CPS was conceived as a collaborative competence in a problem-solving process that was mainly assessed in the articles through computer-supported environments, often highly constrained. This category was dominated by a quantitative research paradigm in which conclusions were drawn based on the assessment of individuals' cognition as represented by their observed communications and actions. In conclusion, work in this perspective can be critiqued as regards (i) the individual rather than group level focus, and (ii) the lack of 'authenticity' in highly constrained simulation environments. For example, it has been questioned whether computer-simulated partners can be designed to reliably mimic realistic conversational partners or the extent to which interacting with computer-simulated partners generalises to interacting with human partners (Webb & Gibson, 2015). It should be noted that, the extent to which computer-simulated partners could fully capture the "real-life" collaboration between humans remains an open question, raising, in turn, issues about external validity.

4.5.2.2 CPS practice

A second group of articles (n = 16, 27%) conceived CPS as a type of pedagogic approach providing the social context in which learning took place. Amongst these, CPS was generally defined as a "teaching strategy", "pedagogical practice", and "educational intervention". The aims were to improve learning and to evaluate the outcome of such CPS pedagogic approaches, whether this was students' CPS skills or other skills and attitudes. To achieve that, authors relied mostly on traditional methods of assessment such as pre- and post-tests, and self-report questionnaires for their data collection.

Adopting a socio-cultural viewpoint of learning, the cognitive development of an individual was mediated by the environment, history, culture, and society they lived in, and knowledge was created by individuals working in collaboration with members of the society in collective activity (Vygotsky, 1978). Hence, CPS was hypothesised to be a promising teaching strategy. In contrast to individual problem solving, CPS research was argued to focus on optimising the benefits of social interaction for facilitating the cognitive development of participants (Gu & Cai, 2019). As described by Albert and Kim (2013), during CPS students worked together in small groups scaffolding each other's learning, while working towards achieving a common goal.

It has been argued that CPS is a complex pedagogical approach, and despite its promises, implementing collaborative activities in the classroom presents a series of challenges (Cáceres et al., 2018). Specifically, collaboration does not occur spontaneously, instead it needs to be guided by appropriate scaffolding. A common feature shared among articles in the group was that students needed instructional support and guidance and therefore the CPS activities were highly structured (e.g., Albert & Kim, 2013; Cáceres et al., 2018; Slof et al., 2012). A teaching strategy proposed by Nelson (1999), was adopted by the reviewed articles in this category for task design in online discussion teaching activities and provided guidelines for implementing collaborative activities within authentic collaborative learning environments. This teaching strategy was argued to combine collaborative learning and problem-based learning strategies, encouraging students to learn by doing (P.-C. Lin et al., 2020).

To provide more structure in the collaborative activity, nine (of 14 empirical) articles utilised technology-supported tools as scaffolding. For example, Cai et al. (2016) used a collaborative script to structure CPS in a real classroom. This involved a set of instructions designed to structure CPS and was based on Vygotsky's concept of the zone of proximal development. It was argued that these scripts help students learn how to interact, collaborate, and solve problems. Similarly, Caceres et al. (2018), designed an interactive script that explicitly allowed students to build an argument during CPS, and investigated the impact of incorporating the construction of arguments into a collaborative learning activity on students' learning. In another article, scripting was used to structure the complex learning-task by dividing it into a sequence of distinct problem phases (Slof et al., 2012).

Another common feature of the articles was the methods used to evaluate the effect of the relevant interventions on students' learning. Most of the empirical articles (10 out of 14) tested students before and after the CPS activity and compared the mean average of their performance. While this approach treated the group interaction as an external influence, some articles (7 out of 14) also analysed the intervening interaction itself, e.g., constructed representations and communicative activities (Slof et al., 2012). When coding student interactions, the unit of analysis used was not common across articles. Five articles focused on the level of the individual using single utterances, while two articles used episodes or events. Finally, quantitative content analysis was commonly used to present the results of the learning process.

Summarising the articles, this view was dominated by a quantitative research paradigm in which conclusions were drawn mainly based on individuals' pre-and post-test performances, aiming to thereby evaluate the outcome of certain CPS pedagogic approaches. Some issues concerning this approach were raised, for example, about testing the average change within an intervention group, while ignoring the analysis of change made by each subject (Cukurova, Luckin, & Baines, 2018). Furthermore, educational contexts are very complex with several dimensions most of which are very difficult to control when investigating the effect of a CPS intervention (Cukurova, Luckin, & Baines, 2018). Finally, CPS practice focus has been criticised for employing a black box approach that makes it difficult to understand, for example, the reasons why some groups do not collaborate effectively (Dillenbourg et al., 1996; Janssen et al., 2010).

4.5.2.3 CPS interaction

A small group of articles (n = 6, 10%) emphasised how CPS activity took place: for example, how students interacted with each other, or which patterns of interactions occurred during a problem-solving activity. Articles adopting a CPS interaction focus shared similarities with the second group regarding the theories used to conceptualise CPS. Specifically, authors used a socio-cultural viewpoint of learning and conceived CPS as an activity. However, they differed in terms of their focus/aims, which were on students’ interactions during the CPS activity rather than on the effect of CPS on student outcomes. Data collection relied mainly on problem-solving tasks (either face-to-face or computer-simulated) during which students’ interactions were recorded.

This group maintains the view that a focus on interaction is necessary to understand the value of learning in collaboration with peers, and how it might be enhanced. Despite existing research regarding the often positive outcomes of collaboration, articles in this group argued that little is yet known of the interaction, particularly within the context of open-learning situations such as open-ended tasks (e.g., Chan & Clarke, 2017; Kumpulainen & Kaartinen, 2003). Authors following the view proposed by Roschelle and Teasley (1995, p. 70), for example, conceived CPS as: “a coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem”. Specifically, CPS was viewed as taking place in a negotiated and shared conceptual space (i.e., joint problem space), which was constructed and maintained via shared language, situation, and activity. Therefore, collaboration was not solely or essentially viewed as a collection of individual efforts, rather it was seen as a social meaning-making activity interdependent with cognition and social relations (e.g., Ding, 2009; Kumpulainen & Kaartinen, 2003). The level of “activity” is understood as a collective accomplishment that gives meaning to individual acts. The latter are visible in behaviours and inferences about the individual goals, the former is a matter of analysis (e.g., of the mediated, collective products/exchanges).

A common feature of these articles involves taking a sequence of utterances and actions, as students engage in the problem-solving tasks, as the unit of analysis, prioritising in this way the interaction within the group. For example, Chan and Clarke (2017) analysed student interactions in a collaborative activity using events that constituted a social interaction with a single identifiable purpose. In addition, Chang et al. (2017) segmented related utterances from students' discourse into threads to understand how they communicated and acted to solve the problem. They argued that the analysis of each individual utterance without considering the contextual move of related utterances would not produce comprehensive results at the conceptual level.

Measuring the frequencies or categorising students' exchanges was seen as inadequate, since it ignores how contextual information influences the individual response (Ding, 2009). Therefore, half of the articles used sequential analysis to acknowledge that the message is a function of its context and to reveal the sequential pattern of complex activities. Apart from sequential analysis, qualitative content analysis followed by description of representative episodes using quotes to illustrate students' interaction has been also applied for the analysis of data derived from sequences of multiple utterances (e.g., Chan & Clarke, 2017).

In sum, CPS interaction articles were generally focused on the social interactions between group members during the CPS process. Both qualitative and quantitative methods were used, with the sequence of utterances as the unit of analysis. Collaboration was not considered only as an individual effort; rather it is seen as a social meaning-making activity, and an accomplishment of the collective. In conclusion, work in this perspective offers critiques of the previous bodies of work that focus on individuals' competences or the impact of CPS on a broader range of individual learning outcomes and argues for a focus on how the collective can make meaningful progress in problem solving and learning.

4.6 Discussion

The current systematic literature review examined and analysed how CPS has been conceptualised in recent empirical and theoretical educational research, extending existing reviews in the topic (i.e., Cukurova, Luckin, & Baines, 2018; Graesser et al., 2018; Oliveri et al., 2017). These results are discussed below in regards to their implications for educational policy, practice, and research.

The most widely cited definition of CPS, provided by the PISA 2015 CPS framework (OECD, 2017a), limited the concept from a social activity to an individualised competence for the purposes of international comparisons. It is important to highlight that this definition was driven by the perceived needs of policy and was influenced by the PISA assessment culture and its historicity, framed in a modernist view of education as a measurable, individual achievement. In the same way that individuals' knowledge of a content domain was inferred from their responses to traditional test questions, individuals' CPS competence was inferred from their observed behaviours during the problem-solving process (Care, Scoular, et al., 2016). It has been argued that students' CPS skills targeted for assessment need to be teachable, measurable in large-scale assessment, and have behavioural indicators that eventually could be assessed by teachers in a classroom setting (Hesse et al., 2015). However, practitioners might consider small-scale, authentic problem solving, more beneficial for developing competent collaborators.

In addition, this review found many articles about individualised assessment of students' CPS competence, which causes a concern for validity and authenticity, raising in turn a big question for research on this topic: can such individualised (and mostly computer-simulated) activity actually be valid as a measure of CPS competence construct? The challenge of making inferences about CPS skills from single communications and actions led researchers to incorporate highly constrained environments to support traditional analyses, e.g., reliability analysis. For example, approaches such as the use of computer-simulated partners and selection of predefined messages have been adopted (e.g., Herborn et al., 2020; Rosen, 2015). As previously noted, such approaches restrict interaction which must be flexible enough to allow students to invent unanticipated behaviours and/or responses (Çakır et al., 2009). As there are several studies currently attempting to develop assessments for the purpose of measuring CPS competence, it is suggested that future research studies need to critically appraise the constraints and affordances they share.

For instance, it has been argued that, in individualised, computer-simulated assessments, students had no opportunities for lengthy conversations, to negotiate and build on each other's ideas, while personality and emotions of team members were also out of scope (Graesser et al., 2018). Additionally, constraints on choice and sequence of actions were considered to provide less opportunities in capturing the wide range of behaviours implicit in CPS (Scoular et al., 2017). It is, therefore, important to investigate whether (or not) the constructs being assessed via individualised assessments are similar to those that might be assessed in a face-to-face situation (Webb & Gibson, 2015).

4.6.1 Message to policy makers

An important message to policy makers that this article aims to convey is about the dangers of backwash of assessment on pedagogy. Indeed, when CPS is limited from a social activity to an individualised competence assessed via constrained individualised activity, it is important that, those making use of the derived measures understand the limitations of the social constructs being developed (Webb & Gibson, 2015). There are major issues related to consequential validity that need to be considered, and thus educational policy needs to be sensitive to authenticity. One serious concern is the inclusion of individualised assessments of students' CPS competence in the curriculum and for the purposes of high-stakes assessment, especially after the increased attention in the topic following the publication of PISA results. The issue concerns the consequential aspect of validity as discussed by Messick (1995), which depends on the particular use of the assessment and the social consequences of test score interpretation and use. There is a risk that teachers might "teach to the test", which means getting students to practise artificial exercises that deviate from "real life" collaboration, so that they can score high marks. There is also a risk that, if teachers rely exclusively on constrained assessment environments, then students might build up unrealistic expectations of what collaboration is (Rosen, 2015). As Webb and Gibson (2015) argue, there are risks of oversimplifying and failing to understand the limitations of assessments. It is, therefore, important that policy makers are aware of those dangers to prevent assessment starting to drive the curriculum rather than supporting it.

4.6.2 Recommendations for further research

This article advocates for research to move through methodologies that complement studies focusing on CPS as competence. Authors of such articles emphasised individual capacities within collaborative situations, as opposed to the collaborating pair/group. Limited research has been conducted so far utilising group-level indicators when assessing students' CPS skills in combination to individual indicators, and the analysis of interaction processes from different levels of description (reflecting individuals, groups, and communities) is currently lacking (Dillenbourg et al., 1996; Fiore et al., 2010; Stahl, 2006). When the impact of CPS as a teaching approach on learning outcomes was investigated without analysing the social interaction itself (CPS practice), it was argued to ignore the wealth of data made available by the collaborative activities themselves (von Davier & Halpin, 2013). Similarly, such an approach was argued to expose very little the collaboration process taking place between the two points of assessment (Seidouvy & Schindler, 2019).

This article, therefore, argues for a focus on how the collective can make meaningful progress in assessing problem solving and learning. Correspondence between the conceptualisation of CPS and the unit of analysis employed to operationalise the construct is considered necessary for the research community to move towards a conceptual coherence in CPS-related research. Future research studies need to carefully develop their conceptualisations considering existing conceptual frames and in relation to the purpose of research and its epistemology. The research paper in Chapter 6 contributes to this gap in the literature by examining validity evidence of the PISA 2015 CPS competence measure, based on the multidimensional character of CPS competence. Drawing from the PISA 2015 CPS framework (OECD, 2017a), the three core competencies for CPS (i.e., Establishing and maintaining shared understanding, Taking appropriate action to solve the problem, and Establishing and maintaining team organisation) are used to investigate whether there is evidence suggesting that the CPS competence construct, as defined and assessed by the PISA 2015 CPS assessment, should be multi-dimensional.

4.6.3 Limitations

This article reviews only English language publications in peer-reviewed journals. Due to these choices, quality work in other languages and work not published in academic journals may have been omitted. An additional limitation relates to the terminology used in our article search approach which may have led to missing certain publications not using the exact key terms, but still engaging with the concept using different terminology. Future studies could offer a more extended analysis of the remaining literature. Nevertheless, the comprehensive overview presented here can serve as a starting point for future reviews and for those who are new to CPS research.

4.7 Conclusion

This article extends current understanding and contributes to knowledge about the conceptualisation of CPS in educational research by analysing definitions and their relation to the methods and units of analysis used in 59 articles. The review identified the strengths and weaknesses of CPS-competence, CPS-practice and CPS-interaction concepts and models and argued that future research should ensure to situate their work in regards to these existing categories to maintain conceptual coherence. It is suggested that the coherent use of such conceptualisations will help overcome previous problems in CPS research lacking in coherence and consistency and help research become more cumulative for the research field itself, but also for policy and practice. Furthermore, this article argues that more information about the validity and authenticity of the individualised assessments of CPS competence is needed. The research paper presented in Chapter 7 contributes to the investigation of authenticity in individualised CPS assessments. Specifically, it examines what a small sample of secondary school students in England say about how they comprehend the items included in the PISA 2015 CPS assessment and how they explain their answers to them through cognitive interviewing. Finally, this article advocates future research studies to focus on the exploration of CPS from different levels of description, extending existing conceptual frames and in relation to the purpose of research and its epistemology.

4.8 References

4.8.1 Full reference list of 59 articles in the review (in alphabetical order)

Albert, L., & Kim, R. (2013). Developing Creativity Through Collaborative Problem Solving. *Journal of Mathematics Education at Teachers College*, *4*(Fall-Winter), 32–38.

Andrews-Todd, J., & Forsyth, C. M. (2020). Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Computers in Human Behavior*, *104*, 105759. https://doi.org/10.1016/j.chb.2018.10.025

Andrews-Todd, J., Jackson, G. T., & Kurzum, C. (2019). Collaborative Problem Solving Assessment in an Online Mathematics Task. *ETS Research Report Series*, *2019*(1), 1–7. https://doi.org/10.1002/ets2.12260

Andrews-Todd, J., & Kerr, D. (2019). Application of Ontologies for Assessing Collaborative Problem Solving Skills. *International Journal of Testing*, *19*(2), 172–187. https://doi.org/10.1080/15305058.2019.1573823

Cáceres, M., Nussbaum, M., Marroquín, M., Gleisner, S., & Marquínez, J. T. (2018). Building arguments: Key to collaborative scaffolding. *Interactive Learning Environments*, *26*(3), 355–371. https://doi.org/10.1080/10494820.2017.1333010

Cai, H., Lin, L., & Gu, X. (2016). Using a semantic diagram to structure a collaborative problem solving process in the classroom. *Educational Technology Research and Development*, *64*(6), 1207–1225. https://doi.org/10.1007/s11423-016-9445-6

Camacho-Morles, J., Slemp, G. R., Oades, L. G., Morrish, L., & Scoular, C. (2019). The role of achievement emotions in the collaborative problem-solving performance of adolescents. *Learning and Individual Differences*, *70*, 169–181. https://doi.org/10.1016/j.lindif.2019.02.005

Care, E., & Griffin, P. (2014). An Approach to Assessment of Collaborative Problem Solving. *Research and Practice in Technology Enhanced Learning*, *9*(3), 367–388.

Care, E., Scoular, C., & Griffin, P. (2016). Assessment of Collaborative Problem Solving in Education Environments. *Applied Measurement in Education*, *29*(4), 250–264. https://doi.org/10.1080/08957347.2016.1209204

Chan, M. C. E., & Clarke, D. (2017). Structured affordances in the use of open-ended tasks to facilitate collaborative problem solving. *ZDM - Mathematics Education*, *49*(6), 951–963. https://doi.org/10.1007/s11858-017-0876-2

Chang, C.-J., Chang, M.-H., Chiu, B.-C., Liu, C.-C., Fan Chiang, S.-H., Wen, C.-T., Hwang, F.-K., Wu, Y.-T., Chao, P.-Y., Lai, C.-H., Wu, S.-W., Chang, C.-K., & Chen, W. (2017). An analysis of student collaborative problem solving activities mediated by collaborative simulations. *Computers and Education*, *114*(C), 222–235. https://doi.org/10.1016/j.compedu.2017.07.008

Chang, C.-J., Chang, M.-H., Liu, C.-C., Chiu, B.-C., Fan Chiang, S.-H., Wen, C.-T., Hwang, F.-K., Chao, P.-Y., Chen, Y.-L., & Chai, C.-S. (2017). An analysis of collaborative problem-solving activities mediated by individual-based and collaborative computer simulations. *Journal of Computer Assisted Learning*, *33*(6), 649–662. https://doi.org/10.1111/jcal.12208

Chao, J., Liu, C.-H., & Yeh, Y.-H. (2018). Analysis of the Learning Effectiveness of Atayal Culture CPS Spatial Concept Course on Indigenous Students. *Eurasia Journal of Mathematics, Science and Technology Education*, *14*(6), 2059–2066. https://doi.org/10.29333/ejmste/86162

Chen, L., Inoue, K., Goda, Y., Okubo, F., Taniguchi, Y., Oi, M., Konomi, S., Ogata, H., & Yamada, M. (2020). Exploring Factors that Influence Collaborative Problem Solving Awareness in Science Education. *Technology, Knowledge and Learning*, *25*(2), 337–366. https://doi.org/10.1007/s10758-020-09436-8

Cho, Y. H., & Lim, K. Y. T. (2017). Effectiveness of collaborative learning with 3D virtual worlds: Collaborative learning. *British Journal of Educational Technology*, *48*(1), 202–211. https://doi.org/10.1111/bjet.12356

Cukurova, M., Luckin, R., & Baines, E. (2018). The significance of context for the emergence and implementation of research evidence: The case of collaborative problem-solving. *Oxford Review of Education*, *44*(3), 322–337. https://doi.org/10.1080/03054985.2017.1389713

Cukurova, M., Luckin, R., Millán, E., & Mavrikis, M. (2018). The NISPI framework: Analysing collaborative problem-solving from students’ physical interactions. *Computers & Education*, *116*, 93–109. https://doi.org/10.1016/j.compedu.2017.08.007

De Boeck, P., & Scalise, K. (2019). Collaborative Problem Solving: Processing Actions, Time, and Performance. *Frontiers in Psychology*, *10*, 1280. https://doi.org/10.3389/fpsyg.2019.01280

Ding, N. (2009). Visualizing the sequential process of knowledge elaboration in computer-supported collaborative problem solving. *Computers & Education*, *52*(2), 509–519. https://doi.org/10.1016/j.compedu.2008.10.009

Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the Science of Collaborative Problem Solving. *Psychological Science in the Public Interest*, *19*(2), 59–92. https://doi.org/10.1177/1529100618808244

Gu, X., & Cai, H. (2019). How a semantic diagram tool influences transaction costs during collaborative problem solving. *Journal of Computer Assisted Learning*, *35*(1), 23–33. https://doi.org/10.1111/jcal.12307

Häkkinen, P., Järvelä, S., Mäkitalo-Siegl, K., Ahonen, A., Näykki, P., & Valtonen, T. (2017). Preparing teacher-students for twenty-first-century learning practices (PREP 21): A framework for enhancing collaborative problem-solving and strategic learning skills. *Teachers and Teaching*, *23*(1), 25–41. https://doi.org/10.1080/13540602.2016.1203772

Harding, S.-M. E., & Griffin, P. (2016). Rasch Measurement of Collaborative Problem Solving in an Online Environment. *Journal of Applied Measurement*, *1*(17), 35–53.

Harding, S.-M. E., Griffin, P., Awwal, N., Alom, B. M., & Scoular, C. (2017). Measuring Collaborative Problem Solving Using Mathematics-Based Tasks. *AERA Open*, *3*(3), 1–19. https://doi.org/10.1177/2332858417728046

Herborn, K., Mustafić, M., & Greiff, S. (2017). Mapping an Experiment-Based Assessment of Collaborative Behavior Onto Collaborative Problem Solving in PISA 2015: A Cluster Analysis Approach for Collaborator Profiles. *Journal of Educational Measurement*, *54*(1), 103–122. https://doi.org/10.1111/jedm.12135

Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? *Computers in Human Behavior*, *104*, 105624. https://doi.org/10.1016/j.chb.2018.07.035

Herro, D., Quigley, C., Andrews, J., & Delacruz, G. (2017). Co-Measure: Developing an assessment for student collaboration in STEAM activities. *International Journal of STEM Education*, *4*(1). https://doi.org/10.1186/s40594-017-0094-z

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

Huang, C. S. J., Su, A. Y. S., Yang, S. J. H., & Liou, H.-H. (2017). A collaborative digital pen learning approach to improving students' learning achievement and motivation in mathematics courses. *Computers & Education*, *107*, 31–44. https://doi.org/10.1016/j.compedu.2016.12.014

Karabulut-Ilgu, A., Yao, S., Savolainen, P., & Jahren, C. (2018). Student Perspectives on the Flipped-Classroom Approach and Collaborative Problem-Solving Process. *Journal of Educational Computing Research*, *56*(4), 513–537. https://doi.org/10.1177/0735633117715033

Krkovic, K., Wüstenberg, S., & Greiff, S. (2016). Assessing Collaborative Behavior in Students: An Experiment-Based Assessment Approach. *European Journal of Psychological Assessment*, *32*(1), 52–60. https://doi.org/10.1027/1015-5759/a000329

Kumpulainen, K., & Kaartinen, S. (2003). The Interpersonal Dynamics of Collaborative Reasoning in Peer Interactive Dyads. *The Journal of Experimental Education*, *71*(4), 333–370. https://doi.org/10.1080/00220970309602069

Li, C.-H., & Liu, Z.-Y. (2017). Collaborative Problem-Solving Behavior of 15-Year-Old Taiwanese Students in Science Education. *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(10), 6677–6695. https://doi.org/10.12973/ejmste/78189

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

Lin, P.-C., Hou, H.-T., & Chang, K.-E. (2020). The development of a collaborative problem solving environment that integrates a scaffolding mind tool and simulation-based learning: An analysis of learners' performance and their cognitive process in discussion. *Interactive Learning Environments*, 1–18. https://doi.org/10.1080/10494820.2020.1719163

Merrill, M. D. (2002). First principles of instruction. *Educational Technology Research and Development*, *50*(3), 43–59. https://doi.org/10.1007/BF02505024

Nouri, J., Åkerfeldt, A., Fors, U., & Selander, S. (2017). Assessing Collaborative Problem Solving Skills in Technology-Enhanced Learning Environments – The PISA Framework and Modes of Communication. *International Journal of Emerging Technologies in Learning (IJET)*, *12*(04), 163. https://doi.org/10.3991/ijet.v12i04.6737

Oliveri, M. E., Lawless, R., & Mislevy, R. J. (2019). Using Evidence-Centered Design to Support the Development of Culturally and Linguistically Sensitive Collaborative Problem-Solving Assessments. *International Journal of Testing*, *19*(3), 270–300. https://doi.org/10.1080/15305058.2018.1543308

Oliveri, M. E., Lawless, R., & Molloy, H. (2017). A Literature Review on Collaborative Problem Solving for College and Workforce Readiness. *ETS Research Report Series*, *2017*(1), 1–27. https://doi.org/10.1002/ets2.12133

O'Neil, H. F., Chuang, S.-H. (sabrina), & Chung, G. K. W. K. (2003). Issues in the Computer-based Assessment of Collaborative Problem Solving. *Assessment in Education: Principles, Policy & Practice*, *10*(3), 361–373. https://doi.org/10.1080/0969594032000148190

Polyak, S. T., von Davier, A., & Peterschmidt, K. (2017). Computational Psychometrics for the Measurement of Collaborative Problem Solving Skills. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.02029

Rosen, Y. (2014). Comparability of Conflict Opportunities in Human-to-Human and Human-to-Agent Online Collaborative Problem Solving. *Technology, Knowledge and Learning*, *19*, 147–164. https://doi.org/10.1007/s10758-014-9229-1

Rosen, Y. (2015). Computer-based Assessment of Collaborative Problem Solving: Exploring the Feasibility of Human-to-Agent Approach. *International Journal* *of Artificial Intelligence in Education*, *25*(3), 380–406. https://doi.org/10.1007/s40593-015-0042-3

Rosen, Y. (2017). Assessing Students in Human-to-Agent Settings to Inform Collaborative Problem-Solving Learning. *Journal of Educational Measurement*, *54*(1), 36–53. https://doi.org/10.1111/jedm.12131

Rosen, Y., & Foltz, P. (2014). Assessing collaborative problem solving through automated technologies. *Research and Practice in Technology Enhanced Learning*, *9*, 389–410.

Rosen, Y., Wolf, I., & Stoeffler, K. (2020). Fostering collaborative problem solving skills in science: The Animalia project. *Computers in Human Behavior*, *104*, 105922. https://doi.org/10.1016/j.chb.2019.02.018

Scoular, C., & Care, E. (2019). A Generalized Scoring Process to Measure Collaborative Problem Solving in Online Environments. *Educational Assessment*, *24*(3), 213–234. https://doi.org/10.1080/10627197.2019.1615372

Scoular, C., & Care, E. (2020). Monitoring patterns of social and cognitive student behaviors in online collaborative problem solving assessments. *Computers in Human Behavior*, 105874. https://doi.org/10.1016/j.chb.2019.01.007

Scoular, C., Care, E., & Hesse, F. W. (2017). Designs for Operationalizing Collaborative Problem Solving for Automated Assessment. *Journal of Educational Measurement*, *54*(1), 12–35. https://doi.org/10.1111/jedm.12130

Siddiq, F., & Scherer, R. (2017). Revealing the processes of students' interaction with a novel collaborative problem solving task: An in-depth analysis of think-aloud protocols. *Computers in Human Behavior*, *76*, 509–525. https://doi.org/10.1016/j.chb.2017.08.007

Slof, B., Erkens, G., & Kirschner, P. A. (2012). The effects of constructing domain-specific representations on coordination processes and learning in a CSCL-environment. *Computers in Human Behavior*, *28*(4), 1478–1489. https://doi.org/10.1016/j.chb.2012.03.011

Slof, B., Erkens, G., Kirschner, P. A., Janssen, J., & Jaspers, J. G. M. (2012). Successfully carrying out complex learning-tasks through guiding teams' qualitative and quantitative reasoning. *Instructional Science*, *40*(3), 623–643. https://doi.org/10.1007/s11251-011-9185-2

Stadler, M., Herborn, K., Mustafić, M., & Greiff, S. (2019). Computer-Based Collaborative Problem Solving in PISA 2015 and the Role of Personality. *Journal of Intelligence*, *7*(3), 15. https://doi.org/10.3390/jintelligence7030015

Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving.

Computers & Education, *143*, 103672. https://doi.org/10.1016/j.compedu.2019.103672

Tawfik, A., Sánchez, L., & Saparova, D. (2014). The Effects of Case Libraries in Supporting Collaborative Problem-Solving in an Online Learning Environment. *Technology, Knowledge and Learning*, *19*(3), 337–358. https://doi.org/10.1007/s10758-014-9230-8

Vista, A., Care, E., & Awwal, N. (2017). Visualising and examining sequential actions as behavioural paths that can be interpreted as markers of complex behaviours. *Computers in Human Behavior*, *76*, 656–671. https://doi.org/10.1016/j.chb.2017.01.027

Webb, M., & Gibson, D. (2015). Technology enhanced assessment in complex collaborative settings. *Education and Information Technologies*, *20*(4), 675–695. https://doi.org/10.1007/s10639-015-9413-5

Wu, S.-Y. (2020). Incorporation of Collaborative Problem Solving and Cognitive Tools to Improve Higher Cognitive Processing in Online Discussion Environments. *Journal of Educational Computing Research*, *58*(1), 249–272. https://doi.org/10.1177/0735633119828044

Yuan, J., Xiao, Y., & Liu, H. (2019). Assessment of Collaborative Problem Solving Based on Process Stream Data: A New Paradigm for Extracting Indicators and Modeling Dyad Data. *Frontiers in Psychology*, *10*, 369. https://doi.org/10.3389/fpsyg.2019.00369

4.8.2 Reference list for Chapter 4

Albert, L., & Kim, R. (2013). Developing Creativity Through Collaborative Problem Solving. *Journal of Mathematics Education at Teachers College*, *4*(Fall-Winter), 32–38.

Alexander, P. A. (2020). Methodological Guidance Paper: The Art and Science of Quality Systematic Reviews. *Review of Educational Research*, *90*(1), 6–23. https://doi.org/10.3102/0034654319854352

Andrews-Todd, J., & Kerr, D. (2019). Application of Ontologies for Assessing Collaborative Problem Solving Skills. *International Journal of Testing*, *19*(2), 172–187. https://doi.org/10.1080/15305058.2019.1573823

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). Springer Netherlands.

Bourdieu, P. (1977). *Outline of a theory of practice*. University Press.

Cáceres, M., Nussbaum, M., Marroquín, M., Gleisner, S., & Marquínez, J. T. (2018). Building arguments: Key to collaborative scaffolding. *Interactive Learning*

Environments, *26*(3), 355–371. https://doi.org/10.1080/10494820.2017.1333010

Cai, H., Lin, L., & Gu, X. (2016). Using a semantic diagram to structure a collaborative problem solving process in the classroom. *Educational Technology Research and Development*, *64*(6), 1207–1225. https://doi.org/10.1007/s11423-016-9445-6

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

Care, E., Anderson, K., & Kim, H. (2016). *Visualizing the Breadth of Skills Movement Across Education Systems*. Center for Universal Education at the Brookings Institution.

Care, E., & Griffin, P. (2014). An Approach to Assessment of Collaborative Problem Solving. *Research and Practice in Technology Enhanced Learning*, *9*(3), 367–388.

Care, E., Scoular, C., & Griffin, P. (2016). Assessment of Collaborative Problem Solving in Education Environments. *Applied Measurement in Education*, *29*(4), 250–264. https://doi.org/10.1080/08957347.2016.1209204

Chan, M. C. E., & Clarke, D. (2017). Structured affordances in the use of open-ended tasks to facilitate collaborative problem solving. *ZDM - Mathematics Education*, *49*(6), 951–963. https://doi.org/10.1007/s11858-017-0876-2

Chang, C.-J., Chang, M.-H., Chiu, B.-C., Liu, C.-C., Fan Chiang, S.-H., Wen, C.-T., Hwang, F.-K., Wu, Y.-T., Chao, P.-Y., Lai, C.-H., Wu, S.-W., Chang, C.-K., & Chen, W. (2017). An analysis of student collaborative problem solving activities mediated by collaborative simulations. *Computers and Education*, *114*(C), 222–235. https://doi.org/10.1016/j.compedu.2017.07.008

Cobb, P., & Gravemeijer, K. (2008). Experimenting to Support and Understand Learning Processes. In A. E. Kelly, R. A. Lesh, & J. Y. Baek (Eds.), *Handbook of Design Research Methods in Education*. Routledge.

Cohen, E. G. (1994). Restructuring the Classroom: Conditions for Productive Small Groups. *Review of Educational Research*, *64*(1), 1–35. https://doi.org/10.3102/00346543064001001

Cukurova, M., Luckin, R., & Baines, E. (2018). The significance of context for the emergence and implementation of research evidence: The case of collaborative problem-solving. *Oxford Review of Education*, *44*(3), 322–337. https://doi.org/10.1080/03054985.2017.1389713

Cukurova, M., Luckin, R., Millán, E., & Mavrikis, M. (2018). The NISPI framework: Analysing collaborative problem-solving from students’ physical interactions. *Computers & Education*, *116*, 93–109. https://doi.org/10.1016/j.compedu.2017.08.007

Dewey, J. (1910). *How we think* (pp. vi, 228). D C Heath. https://doi.org/10.1037/10903-000

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process* (revised edition). D.C. Heath and company.

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

Dillenbourg, P., Baker, M. J., Blaye, A., & O'Malley, C. (1996). The evolution of research on collaborative learning. In E. Spada & P. Reiman (Eds.), *Learning in Humans and Machine: Towards an interdisciplinary learning science* (pp. 189–211). Elsevier.

Ding, N. (2009). Visualizing the sequential process of knowledge elaboration in computer-supported collaborative problem solving. *Computers & Education*, *52*(2), 509–519. https://doi.org/10.1016/j.compedu.2008.10.009

Dominowski, R. L., & Bourne, L. E. (1994). History of Research on Thinking and Problem Solving. In R. J. Sternberg (Ed.), *Thinking and Problem Solving* (Vol. 2, pp. 1–35). Academic Press.

Engeström, Y. (1999). Activity Theory and Individual and Social Transformation. In Y. Engeström, R. Miettinen, & R.-L. Punamäki-Gitai (Eds.), *Perspectives on Activity Theory* (pp. 19–38). Cambridge University Press.

Farra, H. (1988). The Reflective Thought Process: John Dewey Re-visited. *The Journal of Creative Behavior*, *22*(1), 1–8. https://doi.org/10.1002/j.2162-6057.1988.tb01338.x

Fiore, S. M., Rosen, M. A., Smith-Jentsch, K. A., Salas, E., Letsky, M., & Warner, N. (2010). Toward an understanding of macrocognition in teams: Predicting processes in complex collaborative contexts. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(2), 203–224. https://doi.org/10.1177/0018720810369807

Gough, D., Oliver, S., & Thomas, J. (2016). *An introduction to systematic reviews* (2nd ed.). SAGE Publications Ltd.

Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the Science of Collaborative Problem Solving. *Psychological Science in the Public Interest*, *19*(2), 59–92. https://doi.org/10.1177/1529100618808244

Griffin, P., & Care, E. (2014). *Assessment and teaching of 21st century skills: Methods and approach*. Springer.

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

Gu, X., & Cai, H. (2019). How a semantic diagram tool influences transaction costs during collaborative problem solving. *Journal of Computer Assisted Learning*, *35*(1), 23–33. https://doi.org/10.1111/jcal.12307

Harding, S.-M. E., Griffin, P., Awwal, N., Alom, B. M., & Scoular, C. (2017). Measuring Collaborative Problem Solving Using Mathematics-Based Tasks. *AERA Open*, *3*(3), 1–19. https://doi.org/10.1177/2332858417728046

Herborn, K., Mustafić, M., & Greiff, S. (2017). Mapping an Experiment-Based Assessment of Collaborative Behavior Onto Collaborative Problem Solving in PISA 2015: A Cluster Analysis Approach for Collaborator Profiles. *Journal of Educational Measurement*, *54*(1), 103–122. https://doi.org/10.1111/jedm.12135

Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? *Computers in Human Behavior*, *104*, 105624. https://doi.org/10.1016/j.chb.2018.07.035

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A Framework for Teachable Collaborative Problem Solving Skills. In P. Griffin & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 37–56). Springer Netherlands. https://doi.org/10.1007/978-94-017-9395-7_2

Janssen, J., Kirschner, F., Erkens, G., Kirschner, P. A., & Paas, F. (2010). Making the Black Box of Collaborative Learning Transparent: Combining Process-Oriented and Cognitive Load Approaches. *Educational Psychology Review*, *22*(2), 139–154. https://doi.org/10.1007/s10648-010-9131-x

Kumpulainen, K., & Kaartinen, S. (2003). The Interpersonal Dynamics of Collaborative Reasoning in Peer Interactive Dyads. *The Journal of Experimental Education*, *71*(4), 333–370. https://doi.org/10.1080/00220970309602069

Lai, E. R. (2011). *Collaboration: A literature review*. Pearson.

Lave, J. (1991). Situating Learning in Communities of Practice. In L. Resnick, J. Levine, & S. Teasley (Eds.), *Perspectives on Socially Shared Cognition* (pp. 63–83). American Psychological Association.

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

Lin, P.-C., Hou, H.-T., & Chang, K.-E. (2020). The development of a collaborative problem solving environment that integrates a scaffolding mind tool and simulation-based learning: An analysis of learners' performance and their cognitive process in discussion. *Interactive Learning Environments*, 1–18. https://doi.org/10.1080/10494820.2020.1719163

Mayer, R. E. (1992). *Thinking, problem solving, cognition* (2nd ed.). Freeman.

Mayer, R. E. (2009). Information Processing. In T. L. Good (Ed.), *21st Century Education: A Reference Handbook* (pp. 168–174). SAGE Publications, Inc. https://doi.org/10.4135/9781412964012

Mayer, R. E. (2013). Problem Solving. In D. Reisberg (Ed.), *The Oxford Handbook of Cognitive Psychology* (pp. 769–778). Oxford University Press.

Mayer, R. E., & Wittrock, M. C. (2006). Problem Solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology* (2nd ed., pp. 287–304). Erlbaum.

Messick, S. (1995). Validity of Psychological Assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Annals of Internal Medicine*, *151*(4), 264–269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

Nelson, L. M. (1999). Collaborative problem solving. In *Instructional design theories and models: A new paradigm of instructional theory* (Vol. 1–2, pp. 241–267). Lawrence Erlbaum Associates.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hall.

O'Donnell, A., & Hmelo-Silver, C. E. (2013). Introduction: What is Collaborative Learning? : An Overview. In *The international handbook of collaborative learning* (pp. 1–15). Routledge. https://doi.org/10.4324/9780203837290-1

OECD. (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. OECD. https://doi.org/10.1787/9789264190511-en

OECD. (2017a). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. OECD Publishing. https://doi.org/10.1787/9789264281820-en

OECD. (2017b). *PISA 2015 Results (Volume V): Collaborative problem solving*. OECD Publishing. https://doi.org/10.1787/9789264285521-en

Oliveri, M. E., Lawless, R., & Molloy, H. (2017). A Literature Review on Collaborative Problem Solving for College and Workforce Readiness. *ETS Research Report Series*, *2017*(1), 1–27. https://doi.org/10.1002/ets2.12133

O'Neil, H. F., Chuang, S.-H. (sabrina), & Chung, G. K. W. K. (2003). Issues in the Computer-based Assessment of Collaborative Problem Solving. *Assessment in Education: Principles, Policy & Practice*, *10*(3), 361–373. https://doi.org/10.1080/0969594032000148190

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell.

Pólya, G. (1945). *How to solve it: A new aspect of mathematical method*. Princeton University Press.

Polyak, S. T., von Davier, A., & Peterschmidt, K. (2017). Computational Psychometrics for the Measurement of Collaborative Problem Solving Skills. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.02029

Roschelle, J., & Teasley, S. D. (1995). The Construction of Shared Knowledge in Collaborative Problem Solving. In *Computer Supported Collaborative Learning* (pp. 69–97). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-85098-1_5

Rosen, Y. (2015). Computer-based Assessment of Collaborative Problem Solving: Exploring the Feasibility of Human-to-Agent Approach. *International Journal of Artificial Intelligence in Education*, *25*(3), 380–406. https://doi.org/10.1007/s40593-015-0042-3

Rosen, Y., & Foltz, P. (2014). Assessing collaborative problem solving through automated technologies. *Research and Practice in Technology Enhanced Learning*, *9*, 389–410.

Roseth, C. J., Johnson, D. W., & Johnson, R. T. (2008). Promoting early adolescents' achievement and peer relationships: The effects of cooperative, competitive, and individualistic goal structures. *Psychological Bulletin*, *134*(2), 223–246. https://doi.org/10.1037/0033-2909.134.2.223

Schoenfeld, A. H. (1985). *Mathematical problem solving*. Academic Press.

Schoenfeld, A. H. (1987). Pólya, Problem Solving, and Education. *Mathematics Magazine*, *60*(5), 283–291. https://doi.org/10.1080/0025570X.1987.11977325

Scoular, C., & Care, E. (2019). A Generalized Scoring Process to Measure Collaborative Problem Solving in Online Environments. *Educational Assessment*, *24*(3), 213–234. https://doi.org/10.1080/10627197.2019.1615372

Scoular, C., & Care, E. (2020). Monitoring patterns of social and cognitive student behaviors in online collaborative problem solving assessments. *Computers in Human Behavior*, 105874. https://doi.org/10.1016/j.chb.2019.01.007

Scoular, C., Care, E., & Hesse, F. W. (2017). Designs for Operationalizing Collaborative Problem Solving for Automated Assessment. *Journal of Educational Measurement*, *54*(1), 12–35. https://doi.org/10.1111/jedm.12130

Seidouvy, A., & Schindler, M. (2019). An inferentialist account of students' collaboration in mathematics education. *Mathematics Education Research Journal*. https://doi.org/10.1007/s13394-019-00267-0

Slavin, R. E. (1980). Cooperative Learning. *Review of Educational Research*, *50*(2), 315–342. https://doi.org/10.3102/00346543050002315

Slof, B., Erkens, G., Kirschner, P. A., Janssen, J., & Jaspers, J. G. M. (2012). Successfully carrying out complex learning-tasks through guiding teams' qualitative and quantitative reasoning. *Instructional Science*, *40*(3), 623–643. https://doi.org/10.1007/s11251-011-9185-2

Stahl, G. (2006). *Group Cognition: Computer Support for Building Collaborative Knowledge*. The MIT Press.

Stahl, G. (2013). Theories of Cognition in Collaborative Learning. In C. E. Hmelo-Silver, C. A. Chinn, C. Chan, & A. M. O'Donnell (Eds.), *The International Handbook of Collaborative Learning*. Routledge. https://doi.org/10.4324/9780203837290.ch4

Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education*, *143*, 103672. https://doi.org/10.1016/j.compedu.2019.103672

Thomas, J., Brunton, J., & Graziosi, S. (2010). *EPPI-Reviewer 4: Software for research synthesis*. Social Science Research Unit, UCL Institute of Education.

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, *8*(1), 45. https://doi.org/10.1186/1471-2288-8-45

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. Macmillan Press. https://doi.org/10.5962/bhl.title.55072

Vista, A., Care, E., & Awwal, N. (2017). Visualising and examining sequential actions as behavioural paths that can be interpreted as markers of complex behaviours. *Computers in Human Behavior*, *76*, 656–671. https://doi.org/10.1016/j.chb.2017.01.027

von Davier, A., & Halpin, P. (2013). Collaborative Problem Solving and the Assessment of Cognitive Skills: Psychometric Considerations. *ETS Research Report Series*, *2013*(2), i–36. https://doi.org/10.1002/j.2333-8504.2013.tb02348.x

Vygotsky, L. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Webb, M., & Gibson, D. (2015). Technology enhanced assessment in complex collaborative settings. *Education and Information Technologies*, *20*(4), 675–695. https://doi.org/10.1007/s10639-015-9413-5

Yuan, J., Xiao, Y., & Liu, H. (2019). Assessment of Collaborative Problem Solving Based on Process Stream Data: A New Paradigm for Extracting Indicators and Modeling Dyad Data. *Frontiers in Psychology*, *10*, 369. https://doi.org/10.3389/fpsyg.2019.00369

Chapter 5 Assessment of students’ collaborative problem-solving competence with the use of computer-simulated, scenario-based tasks: A systematic literature review

5.1 Abstract

Assessment of collaborative problem solving (CPS) competence has recently seen much research interest in education, although the nature of the construct is not well understood yet. Specifically, there is an increased interest in using computer-simulated, scenario-based assessment tasks to provide environments for students to work together with the aim of solving problem scenarios. However, issues have been raised about the validity and authenticity of such collaborative tasks being used for the assessment of students' CPS competence within existing research. As a result, this review provides an analysis of relevant, and systematically selected, empirical articles assessing students' CPS competence with the use of computer-simulated, scenario-based tasks. It aims to: i) describe the characteristics of the assessments, ii) present a summary of the facets of CPS competence measured, and iii) evaluate the strategies adopted for validating the CPS competence measures. This review includes 26 articles reporting 15 assessments. The results indicate that most of the assessments target secondary students working in groups of two and communicating through text messages. In addition, assessments measuring active listening, audience awareness, team empowerment, and team learning, are limited. Further, this review demonstrates that evidence concerning external, substantive, and consequential validity aspects is lacking. Based on these findings, recommendations for future research were made with the aim to develop more comprehensive assessment instruments and validate existing CPS competence measures.

Keywords: collaborative problem solving (CPS), assessment, measurement, skills, validity, systematic literature review

5.2 Introduction

As a competence, CPS is claimed to be necessary for students to succeed in group problem solving activities in today's education and future employment environments. The increasing interest in this concept is also highlighted by the growing number of initiatives to develop assessments that measure students' CPS competence, and consequently provide evidence that students are equipped with skills to meet the CPS demands of their future careers. For example, intergovernmental economic organisations such as the Organisation for Economic Cooperation and Development (OECD) developed a large-scale approach towards measuring students' CPS competence as part of its Programme for International Student Assessment (PISA) 2015 study to inform education systems and policy makers to develop programmes that would improve students' collaboration skills (OECD, 2017b). At a national level, the US National Centre for Education Statistics has recently considered adding an assessment of students' CPS competence as part of the largest nationally representative assessment, i.e., National Assessment of Educational Progress (Fiore et al., 2017). Finally, the timeliness of the topic is highlighted by the recent publication of special issues in journals, such as Applied Measurement in Education (Greiff & Kyllonen, 2016a), Journal of Educational Measurement (von Davier, 2017) and Computers in Human Behavior (Graesser et al., 2020), specifically targeting developments in the assessment and measurement of CPS competence.

To evaluate whether students are equipped with this complex and multi-faceted CPS competence, new assessments have been recently developed (Andrews-Todd & Forsyth, 2020; Griffin & Care, 2014; von Davier et al., 2017). However, an absence of consensus on how to operationalise CPS competence, makes the development and evaluation of assessments challenging. Traditional assessment types such as paper-and-pencil, multiple-choice tests are considered inappropriate for capturing the complexity of CPS competence (Andrews-Todd & Forsyth, 2020; Care & Griffin, 2014). As a result, recent developments in computer simulations are now being implemented to develop scenario-based assessment tasks that capture students' actions and discourse as they engage with a task. There is, therefore, a need to synthesise existing knowledge about these assessment instruments as well as to evaluate the validity evidence for the derived CPS competence measures.

Two recent literature reviews on CPS competence focus primarily on the definition of the concept (Graesser et al., 2018; Oliveri et al., 2017). The first review (Oliveri et al., 2017) presents various theoretical frameworks of CPS competence, limited to higher education and workplace contexts. Although valuable, this review does not include assessments comprising of computer-simulated, scenario-based tasks. In the second review (Graesser et al., 2018), two CPS frameworks for secondary school students are presented, followed by examples of technological advances relevant to the assessment of CPS competence. Nevertheless, the process of identifying and including studies in the review is not made explicit (Gough et al., 2016). In fact, an explicit perspective on the assessment of CPS competence across educational settings with the use of computer-simulated, scenario-based tasks is still lacking.

In the recent CPS concepts review (Chapter 4), the literature on CPS was systematically selected and categorised, focusing on the CPS “concept” in research practice over the last almost two decades, across educational settings. Using this as a backdrop (as will be detailed next), the present article takes a closer look at the methods of data collection (i.e., assessment instruments) and analysis of articles assessing students’ CPS competence with the use of computer-simulated, scenario-based tasks, to develop a better understanding and a critique of existing assessments of CPS competence. Following a systematic literature review methodology (Gough et al., 2016; Petticrew & Roberts, 2006), this article contributes to knowledge about the assessment of students’ CPS competence via three aims: i) to describe the characteristics of the existing CPS assessments, ii) to categorise the facets of CPS competence targeted for measurement, and iii) to evaluate the strategies adopted for validating the CPS competence measures.

The next section presents the background including the research questions, followed by an overview of the systematic review methodology adopted here. The article continues with the results, followed by a discussion and conclusion.

5.3 Background

5.3.1 Conceptualisations of collaborative problem solving

The recent CPS concepts review (Chapter 4) systematically examined the way that CPS has been conceptualised in empirical and theoretical educational research. A total of 59 articles were deemed relevant for review and the analysis led to three categories of conceptualisations: CPS competence, CPS practice, and CPS interaction. Articles adopting CPS competence (n = 36) emphasised individual capacities within collaborative situations, as opposed to focusing on the collaborating pair/group. They focused on assessing CPS as an intended learning outcome, making assumptions about the individual (student's) CPS competence underlying their observed behaviours. In CPS practice (n = 17), authors generally emphasised the individual cognition, investigating how it was affected by cultural mediations, or even collaborative interactions. CPS was then conceived as a pedagogical approach, providing the social context in which learning could take place. Finally, in CPS interaction (n = 6), CPS was conceived as taking place in a negotiated and shared conceptual space in which individual student contributions could not be distinguished from the group co-constructed meaning making process.

Articles using computer-simulated, scenario-based tasks to measure students' CPS competence were found to adopt (almost) exclusively a CPS competence focus. These articles form the focus of the current article as detailed in the methods section below. Specifically, CPS was defined as a competence that is needed for a student to be an effective contributor to a CPS activity. The PISA 2015 CPS framework (OECD, 2017a) and the framework for teachable CPS skills (Hesse et al., 2015) were mostly used to define CPS competence and/or develop assessment instruments. The purpose of such frameworks is to initially describe a construct of interest with enough specificity (e.g., by detailing its specific components) to then guide assessment design and item development (Wolfe & Smith, 2007a). Developing a detailed description of the components that constitute a construct, is one of the central steps in the assessment development process (Mislevy et al., 2003). Tasks are then designed to elicit specific student actions hypothesised to demonstrate the components outlined in the frameworks.

However, such frameworks have been described as reductive, leading to complex constructs being broken down (or reduced) into simpler, more quantitatively manageable constituent components (Vista et al., 2017). Although widely used within educational measurement contexts, this approach may limit the scope of information captured, and ultimately break down what perhaps is or ought to be a holistic activity. In this sense, such an approach might result in a simplified version of what is otherwise considered a very complex construct. Furthermore, this reductive approach is almost universal in tests of competence/problem solving across the curriculum such as in mathematics and science. The question is: Is a student who can competently perform calculations a competent user of problem solving with arithmetic? It is therefore important to note these limitations in the selection of the articles for review, which are relevant for the analysis and the kind of results that will be presented.

5.3.2 Frameworks of collaborative problem-solving competence

As previously described, CPS competence is analysed based on various components, which are further broken down into skills that could be identified in students' actions. These essentially provide the language and basis for task development. Several theoretical frameworks were found to define CPS competence, following varying approaches to mapping different construct-components (e.g., Hesse et al., 2015; OECD, 2017a; Oliveri et al., 2017). Despite the increase, and variety, in construct models analysing CPS skills, there is still a lack of consensus in the literature regarding what constitutes CPS competence, which, in turn, makes developing assessments challenging.

To map the CPS competence components targeted by existing assessments, this article uses a higher education readiness CPS framework (Oliveri et al., 2017) as a coding template with the aim of maintaining a common language in the field. The framework was developed as part of a literature review on CPS competence in higher education contexts. It aimed to inform the assessment of individuals' CPS competence, and defined CPS as a collection of skills that are arguably important for daily life, work, and schooling in the 21st century. The framework organised skills in four components: teamwork, communication, leadership, and problem solving (for a more in-depth overview, please refer to Oliveri et al., 2017). **Teamwork** includes processes related to promoting team cohesion, team empowerment, team learning, self-management and self-leadership, and attitudes of open-mindedness, adaptability, and flexibility. **Communication** involves active listening and information exchange skills. For the **leadership** component, organising activities and resources, monitoring performances, reorganising when faced with obstacles, resolving conflicts, and demonstrating transformational leadership, are considered relevant skills. Finally, **problem solving** includes processes related to identifying and defining a problem, brainstorming, planning, interpreting and analysing information, and evaluating and implementing solutions.

During the review process, some revisions of the framework had to be made to accommodate a more specific perspective on assessment across educational settings. For that reason, two additional CPS competence frameworks were examined, to ensure that relevant components are included (Hesse et al., 2015; OECD, 2017a). These were selected because of their focus on secondary school students, and since they were found to be widely used by articles assessing CPS competence (Chapter 4). The revised coding tool is detailed in the methods section (Chapter 3).

5.3.3 Computer-simulated, scenario-based tasks

While there may be general agreement that CPS competence is important (Griffin et al., 2012), there are issues with the lack of consensus on how to build a collaborative assessment that accurately measures CPS competence in both face-to-face and computer-based simulation environments (Cukurova, Luckin, Millán, et al., 2018; Sun et al., 2020; von Davier et al., 2017). For example, when assessing CPS competence, a decision must be made about whether to assess the individual within a group, the group on its own, or both, while such a decision has implications for measurement. Another challenge is to identify, and evaluate, procedures such as students working together, discussing the problem, and sharing resources, so that they can be scored in an automatic way (Care & Griffin, 2014). Determining the meaning of students' behaviour may be much more complex because of the dynamics and the volume of data generated in collaborative assessment environments (Fiore et al., 2017).

Computer-simulated, scenario-based tasks provide the opportunity to observe students' behaviours in a problem scenario and draw inferences about their CPS competence (Andrews-Todd & Forsyth, 2020; Griffin & Care, 2014; Hao et al., 2017). Some common task designs have constrained the problem space and communication to help identify evidence of students' skills in large streams of process data. For example, several tasks have incorporated computer-simulated participants and predetermined chat messages, in an attempt to provide more control over the collaborative interaction (Hsieh & O'Neil, 2002; OECD, 2017a; Rosen & Foltz, 2014). However, it has been argued that such task design decisions do not always allow for the full scope of CPS competence construct to be measured (Andrews-Todd & Kerr, 2019). Thus, in many cases, the validity of the CPS competence measures derived from such constrained assessments has been questioned (Rosen & Foltz, 2014; Scoular & Care, 2020). Given the complex nature of CPS competence and the lack of consensus in its operationalisation, it is important to review the characteristics of existing assessments using computer-simulated, scenario-based tasks, and the strategies used for validating the derived measures of CPS competence.

5.3.4 Research questions

Three main points have been highlighted in the outlined background: i) the lack of consensus on the state-of-the-art addressing the assessment of CPS competence with the use of computer-simulated, scenario-based tasks; ii) the current conceptual work detailing the nature and content of CPS competence which points towards a complex and likely multidimensional construct; and iii) the concerns about validity of existing CPS competence measures. In light of these, the current article aims to provide an overview of the literature on CPS competence assessment, especially in relation to the use of computer-simulated, scenario-based tasks.

More specifically, the following research question and sub-questions guide the systematic literature review:

> RQ2. How has students’ CPS competence been assessed using computer-simulated, scenario-based assessment tasks in educational research studies?
>
> > RQ2.a: What are the existing assessments of students’ CPS competence and their characteristics (e.g., subject domain, task design features)?
> >
> > RQ2.b: Which facets of CPS competence do the assessments measure?
> >
> > RQ2.c: What strategies for validating CPS competence measures are reported?

This article adds value to the existing literature and contributes to knowledge about the assessment of CPS competence in three ways. First, it offers a systematic review of the state-of-the-art of CPS competence assessment, especially in relation to the computer-simulation and measurement subfield, that will make a significant contribution in informing future research (including future reviews and meta-analyses). Second, this article offers an evaluation of the relationship between conceptualisation of CPS competence and assessment instruments. By highlighting current facets of CPS competence that are not being measured yet, this article will inform the development of more comprehensive assessment instruments. Third, this article offers an evaluation of the strategies used to validate existing CPS competence measures. By highlighting the types of validity evidence that have been overlooked, this article will facilitate the development of future validation studies.

5.4 Methods

5.4.1 Systematic review methodology

A review of research is a form of research in itself, and a systematic review, like the one herein, is 'systematic' in the same way that any empirical research needs to be systematic and transparent, so that the results can be interpreted and assessed in the light of how they were produced. The steps described by Gough et al. (2016) and Petticrew and Roberts (2006) to guide systematic literature reviews were followed, i.e., developing research questions, determining the types of studies to be located, carrying out a comprehensive literature search, formulating inclusion criteria, appraising study quality, and extracting and synthesising data.

5.4.2 Literature database: article search approach and inclusion criteria

An electronic search was conducted in four scientific databases covering research in social sciences (including educational research): SCOPUS, Web of Science, Education Resources Information Centre (ERIC), and British Education Index (BEI). Assessments of CPS competence were previously reported in the fields of education, business, health and medicine (Oliveri et al., 2017). This review was focussed on the field of education, since students emerging from schools into the workforce and public life are expected to be able to work in teams to solve diverse problems (Rosen & Foltz, 2014). The search was first conducted in January 2019 and was updated in April 2020 and December 2020 to include the most recent publications.

Boolean operators were employed to combine the key terms as follows: "collaborative problem solving" AND (student* OR pupil* OR learner*), making the search specific to student populations. The search terms were applied to the fields of title, abstract, and keywords.

Articles were included based on the following criteria:

- peer-reviewed journal articles,
- published between 2010 and 2020,
- written in the English language,
- full text available,
- reported data collection and analysis (primary or secondary),
- referred to student populations from educational settings (from reception to higher education),
- were concerned about the assessment of students' CPS competence in the field of education,
- provided a clear definition/conceptualisation/framework of CPS competence,
- developed (or used, for secondary data) at least one computer-simulated, scenario-based task as their assessment instrument.

To include timely and up-to date articles, the period between 2010 and 2020 was chosen to map current literature published in the last decade. Peer review was used as a quality check criterion, although it is recognised that this has its own limitations (Alexander, 2020). Additionally, a publication was retained for review when assessment of students' CPS competence formed part of the content and focus of the article, meaning that CPS was not used merely as an example among other learning outcomes or as a term specific to a field other than education. The selection of articles was restricted to those using computer-simulated, scenario-based tasks, excluding articles using solely other task types such as self-assessments or teacher evaluations, for two reasons: (i) a recent review (Oliveri et al., 2017) has already included examples of the above task types with the exception of computer-simulated, scenario-based tasks, and (ii) there is an increasing number of studies using computer-simulated, scenario-based tasks to draw inferences about students' CPS competence that have not been reviewed in a systematic manner to date.

5.4.3 Selection of articles

To screen and select articles to be included in the literature review, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) method (Moher et al., 2009) was used. Publication records were managed in the systematic review software EPPI-Reviewer 4 (Thomas et al., 2010). The search yielded 665 articles in total. First, duplicates were excluded across databases resulting in 371 unique articles to be screened. Screening was conducted in two rounds: 1) applying the inclusion criteria to titles and abstracts, and 2) applying the inclusion criteria to the full text of the remaining articles (Figure 5.1). If a decision for inclusion could not be made by reading only the title and abstract, then the article was screened on full text.

Referential backtracking and researcher checking were used to look beyond the results of the database searches (Alexander, 2020). This targeted search entailed examining the reference lists of articles included in the review. The reference lists of existing literature reviews (i.e., Graesser et al., 2018; Oliveri et al., 2017), were also used for referential backtracking even though those documents fell outside the inclusion criteria. Finally, the publication records of authors, who were found to frequently contribute to the topic, were searched. This resulted in identification of four additional articles, bringing the total to 26 articles for inclusion in this review (Figure 5.1; the full list is included in the references). Randomly chosen articles (n = 40) were reviewed by the second author, resulting in a 95% match in the ratings (i.e., inclusion and exclusion) of articles by both reviewers. Any disagreements were resolved through discussion.

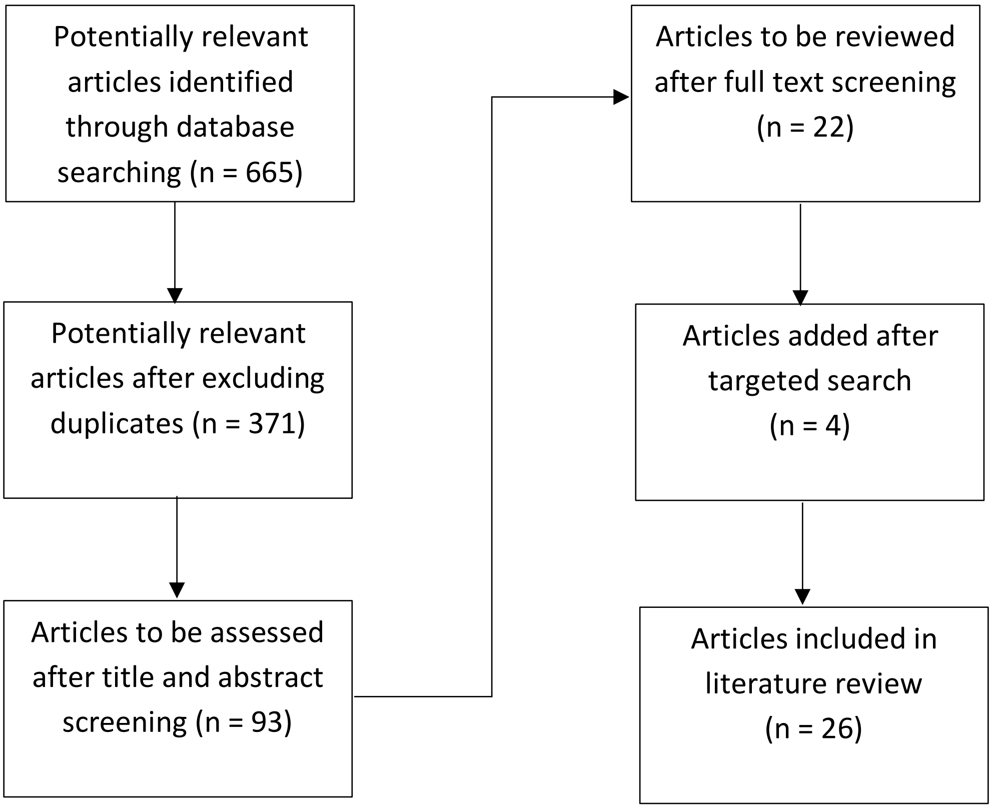

Potentially relevant articles identified through database searching (n = 665)
Potentially relevant articles after excluding duplicates (n = 371)
Articles to be assessed after title and abstract screening (n = 93)
Articles to be reviewed after full text screening (n = 22)
Articles added after targeted search (n = 4)
Articles included in literature review (n = 26)

Figure 5.1. Flow diagram of systematic literature review selection

5.4.4 Coding and data extraction tools

5.4.4.1 Assessment characteristics

To answer RQ2.a, the units of analysis were the assessments, and the information in the articles represented the source of data extracted. In the 26 articles included in this systematic review, 15 distinct assessments of CPS competence were represented. The following format for extracting relevant information about the assessments was developed:

- General information and research design: authors, year of publication, country(ies) of data collection, instruments of data collection, sample size, educational level.
- Task design features: group size, partner mode (i.e., human, or computer-simulated), subject domain, communication mode, scoring approach.

5.4.4.2 Facets of CPS competence

To answer RQ2.b, the units of analysis were the assessments, and the information in the articles were all used as the source of data to inform what facets of CPS competence were targeted by the different assessments. The reporting of the content of each assessment was scrutinised with the aim of identifying the facets, i.e., skills within components of CPS competence, measured by the assessments. Consequently, articles included in the review were appraised against the revised coding tool (a detailed description of the coding tool has been provided in Chapter 3).

5.4.4.3 Evaluation of CPS competence measure

To answer RQ2.c, the units of analysis were the articles, which were coded by extracting information about the strategy for validating CPS measures that they followed. The current article aims to summarise the reported validity evidence in the reviewed articles that undertook some validation work. Following Messick (1995), the six distinct validity aspects emphasising content, substantive, structural, generalisability, external, and consequential aspects of validity, were adopted when coding validity evidence in the reviewed articles (a detailed description of the coding tool has been provided in Chapter 3). Taken together, these aspects provide a way of addressing the multiple and interrelated validity questions that need to be answered in justifying score interpretation and use.

5.5 Results

The selection process resulted in 26 articles to be included in the literature review, which represented 15 assessments of CPS competence targeting students from primary to tertiary educational settings. In the following sections, the results addressing the three research questions are presented.

5.5.1 What are the existing assessments of students' CPS competence and their characteristics?

An overview of the general characteristics of the 15 assessments is provided in Table 5.1.

5.5.1.1 Countries

Table 5.1 shows that most assessments (n = 12) were used to collect data from a single country and only three assessments were administered in at least two countries. In addition, it is worth noting that these three were developed within international projects, i.e., the Assessment and Teaching of 21st Century Skills (Griffin et al., 2012), the PISA 2015 CPS assessment (OECD, 2017b), and the 21st Century Skills Assessment project (Rosen, 2015).

5.5.1.2 Respondents

Sample sizes ranged between 10 and 53,855 students, covering from primary up to higher education level. However, most of the assessments sampled only secondary school students (n = 13). Two assessments sampled higher education students, while only one assessment sampled primary (in addition to secondary) school students.

5.5.1.3 Data collection instruments

Most of the assessments consisted of multiple sets of problem-solving tasks (n = 8), while the remaining used a single problem-solving task. The table shows other instruments used along with the tasks for data collection: self-report questionnaires (n = 6), performance tests in subject domains (n = 3), teacher questionnaires (n = 2), interviews (n = 1), and "think-aloud" protocols (n = 1).

5.5.1.4 Subject domain

Almost half of the assessments (n = 7) included problem-solving tasks that required students to draw on knowledge acquired through traditional subject domains. Six assessments included problem-solving tasks that were claimed to be independent of any subject, which students were expected to solve without having any specialised subject knowledge. Finally, two assessments included a combination of content-free and content-dependent tasks.

5.5.1.5 Group size and composition

The table shows that dyads were mostly used (n = 8), as compared to groups of three (n = 3) and four (n = 1) members. In three assessments, group size was not fixed, instead it ranged from two to four members. In addition, in most of the assessments (n = 9), students were placed in groups composed of human participants only, while five assessments placed students in groups where humans were replaced by computer-simulated partners to facilitate various task design requirements. Finally, one assessment was found to employ both group formats. It should be noted that, questions regarding the extent to which computer-simulated partners could fully capture the real collaboration between humans were raised in the studies (Rosen, 2015; Scoular & Care, 2019).

5.5.1.6 Communication mode

In most assessments (n = 12), communication among the group members was facilitated via written text only. By using free-form text messages generated by students in an embedded chat box, it was claimed that students could interact with each other in an un-scaffolded manner, similar to text messaging, and other forms of media students were familiar with (e.g., Harding et al., 2017). When phrase-chat communication was used, this took the form of a list of pre-defined messages to select from. Communication was treated as a traditional multiple-choice test giving the same set of messages to every student completing the task, and was used mainly in combination with the use of computer-simulated partners (e.g., De Boeck & Scalise, 2019). Only two assessments facilitated communication via audio and video, which allowed students to see and hear each other via webcams and microphones, allowing in this way the observation of tone and body language. This system for communication was claimed to allow students to openly communicate, in an unconstrained way, as they would in many everyday contexts such as in school (Andrews-Todd & Forsyth, 2020). Finally, two assessments used a combination of the above communication approaches.

5.5.1.7 Scoring

Algorithms were programmed to search the recorded actions and chats to identify the specific events reflecting students' skills, allowing for automatic scoring (n = 8). Prior to that, researchers had specified which actions and chats were indicative of specific skills. For example, the presence of any actions on each page of the task was inferred to be indicative of 'participating' (Scoular & Care, 2020). In addition, it is worth noting that, assessments allowing the exchange of free-form messages did not consider the content of those messages when employing automatic scoring. The only exception was identifying the presence of a few keywords (e.g., 'can' and '?' indicating a question) when scoring (e.g., Harding et al., 2017). After identifying the different events and patterns that allowed drawing an inference about students' skills, these were then treated as individual items on a test. When an event or pattern was identified, full credit (or partial credit) was allocated to that student, and when that event or pattern was absent, no credit was allocated.

Alternatively, manual scoring was employed (n = 6), in which students' communicative acts (verbal sentences and utterances) and actions were first segmented into units of analysis and were then assigned codes based on their content. For instance, the communicative act "What do you see on your screen right now?", was coded as indicative of the skill 'discovering abilities and perspective of team members', and each time similar events occurred, a point was accumulated for that code (Nouri et al., 2017). The result of such coding and scoring process was a quantitative frequency description of the occurrence of the different codes.

Table 5.1. Summary descriptions of assessments represented in articles included in the systematic review.

Assessment	Author (year)	Country of data collection	Sample size	Education level	Number of tasks	Subject domain	Group size	Partner mode	Communicati on mode	Scoring	Other methods
1. Assessment and Teaching of 21st Century Skills – CPS tasks	Camacho-Morles et al. (2019)	Australia	200	Secondary	5	CD, CF	2	H-H	Text (free)	Automatic	SQ
	Care and Griffin (2014)	Australia, Costa Rica, the Netherlands, Finland, Singapore, USA	4,056	Secondary	11	CF	2	H-H	Text (free)	Automatic	-
	Harding et al. (2017)	Australia, Costa Rica, the Netherlands, Finland, Singapore, USA	3,004	Secondary	4	CD, CF	2	H-H	Text (free)	Automatic	-
	Harding and Griffin (2016)	Australia, Costa Rica, the Netherlands, Finland, Singapore, USA	3,402	Secondary	4	CD, CF	2	H-H	Text (free)	Automatic	-
	Scoular and Care (2019)	Australia	1,210	Secondary	3	CD, CF	2	H-H	Text (free)	Automatic	-
	Scoular and Care (2020)*	Australia	3,010	Secondary	3	CD, CF	2	H-H	Text (free)	Automatic	-

2. Assessment system for CPS in STEM education	Lin et al. (2015)	Taiwan	222	Secondary	8	CD	2-3	H-A	Text (free)	Insufficient reporting	-
3. Australian Council for Educational Research General Capabilities Assessment	Scoular et al. (2020)*	Australia	1,145	Primary, Secondary	1	CF	3	H-H	Text (free)	Manual	-
4. Circuit Runner CPS game	Polyak et al. (2017)	USA*	159	Secondary	1	CF	2	H-A	Text (phrase)	Automatic	-
5. Collaborative Behaviour Assessment	Herborn et al. (2017)	Germany	481	Secondary	6	CF	2	H-A	Text (free, phrase)	Automatic	SQ, T
	Krkovic et al. (2016)	Germany	483	Secondary	6	CF	2	H-A	Text (free, phrase)	Automatic	SQ
6. Computer-based CPS assessment	Yuan et al. (2019)	China	434	Secondary	5	CF	2	H-H	Text (free)	Automatic	-
7. Computer-based CPS assessment in science	Kuo et al. (2020)	Taiwan	53,855	Secondary	5	CD	2-3	H-A	Text (phrase)	Automatic	-
	Li and Liu (2017)	Taiwan	52,110	Secondary	2	CD	2-3	H-A	Text (phrase)	Automatic	-
8. Computer-based CPS assessment task system	Nouri et al. (2017)	Sweden	24	Secondary	2	CD	2	H-H	Audio, Text (free)	Manual	-
9. CoSketch CPS task	Siddiq and Scherer (2017)	Norway	11	Secondary	1	CD	4	H-H	Text (free)	Manual	P

10. Delivery Hero Assessment	Scoular and Care (2020)*	Australia	3,010	Secondary	3	CD, CF	2	H-H	Text (free)	Automatic	-
	Scoular et al. (2020)*	Australia	1,080	Secondary	3	CD, CF	2	H-H	Text (free)	Automatic	
11. Minecraft Hour of Code CPS task	Sun et al. (2020)	USA	111	Tertiary	1	CD	3	H-H	Video, Audio	Manual	SQ, T
12. Programme for International Student Assessment (PISA) 2015 CPS Assessment	De Boeck and Scalise (2019)	USA	986	Secondary	1	CF	3	H-A	Text (phrase)	Automatic	-
	Herborn et al. (2020)	Germany	386	Secondary	4	CF	2-4	H-A	Text (phrase)	Automatic	-
	Scoular et al. (2020)*	Australia	4,305	Secondary	6	CF	2-4	H-A	Text (phrase)	Automatic	-
	Stadler et al. (2019)	Germany	483	Secondary	4	CF	2-4	H-A	Text (phrase)	Automatic	SQ, T
	Stadler et al. (2020b)	Germany	483	Secondary	5	CF	2-4	H-A	Text (phrase)	Automatic	SQ, TQ, T
13. The Zoo Quest Task	Rosen (2014)	USA, Singapore, Israel	179	Secondary	1	CF	2	H-A, H-H	Text (phrase)	Automatic	-
	Rosen (2015)	USA, Singapore, Israel	179	Secondary	1	CF	2	H-A, H-H	Text (phrase)	Automatic	SQ
	Rosen and Foltz (2014)	USA, Singapore, Israel	179	Secondary	1	CF	2	H-A, H-H	Text (phrase)	Automatic	SQ
14. Three-Resistor Activity	Andrews and Forsyth (2020)	USA*	129	Tertiary	1	CD	3	H-H	Text (free)	Manual	SQ, TQ
15. T-shirt Math Task	Andrews et al. (2019)	USA*	10	Secondary	1	CD	2	H-H	Video, Audio	Manual	I

Notes:

In the category ‘Assessment’, assessments are presented in alphabetical order.

In the category ‘Author’, the asterisk indicates that the article described more than one assessment.

In the category ‘Country of data collection’, the asterisk indicates that the country was not reported, but assumed by the affiliation of the authors.

In the category ‘Subject domain’, the following abbreviations have been used: CD=Content-dependent, CF=Content-free.

In the category ‘Partner mode’, the following abbreviations have been used: H-H=human-to-human approach, H-A=human-to-computer-simulated agent/partner approach.

In the category ‘Other methods’, the following abbreviations have been used: SQ=student questionnaire, TQ=teacher questionnaire, T=subject test, P=think-aloud protocols, I=interview.

5.5.2 Which facets of CPS competence do the assessments measure?

Assessments were associated with either a self-developed or an international framework of CPS competence. It should be noted that no alignment with a national curriculum was reported. As shown in Table 5.2, the most frequently used frameworks were found to be the PISA 2015 CPS framework (OECD, 2017a) and the Framework for teachable CPS skills (Hesse et al., 2015). Appendix 16 provides a description of existing frameworks in more detail.

Table 5.2. Frameworks used by assessments to define collaborative problem solving

Framework	Assessment
Programme for International Student Assessment 2015 CPS framework (OECD, 2017a)	2, 5, 7, 8, 12, 13
Framework for teachable CPS skills (Hesse et al., 2015)	1, 6, 9, 10
CPS ontology (Andrews-Todd & Forsyth, 2020)	14, 15
Australian Council for Educational Research Framework for collaboration (Scoular, Duckworth, et al., 2020)	3
Generalised competency model (Sun et al., 2020)	11
Holistic framework (Camara et al., 2015)	4

Skills (or facets within components) targeted by each one of the assessments are presented in Table 5.3. Several skills related to Teamwork and Communication components, i.e., ‘team empowerment’, ‘team learning’, ‘active listening’, and ‘audience awareness’, were scarcely or not at all covered. Interestingly, the assessments targeting ‘active listening’ and ‘team empowerment’ were the same, suggesting that their specific task characteristics (i.e., video and audio communication, manual scoring) might have allowed for those skills to be captured. Specifically, it can be assumed that it is easier to elicit those skills in environments allowing for more authentic communication. ‘Audience awareness’ relates to students adjusting their contributions to suit other group members’ needs. In constrained communication environments, this is anticipated to be difficult to capture, since students are not able to use non-verbal cues or tone to help them understand other members’ needs.

Table 5.3. Frequency of skills targeted by the assessments

	Assessments														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Component: Teamwork															
Team cohesion	X	X					X	X	X	X		X			
Team empowerment											X			X	X
Team learning															
Self-management and self-leadership	X	X	X	X	X	X	X	X	X	X	X	X		X	X
Open-mindedness, adaptability, and flexibility	X		X			X			X	X	X				
Component: Communication															
Active listening											X			X	X
Exchanging information	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Audience awareness	X														
Component: Leadership															
Organising activities and resources	X	X	X		X	X	X	X	X		X	X		X	X
Reorganising when faced with obstacles		X	X	X			X	X			X	X		X	X
Resolving conflicts	X		X			X			X	X	X			X	X
Component: Problem solving															
Brainstorming and identifying problems	X	X		X	X	X	X	X	X	X	X	X	X	X	X
Interpreting and analysing information	X	X			X	X	X	X	X			X		X	X
Planning and implementing solutions	X	X		X	X	X	X	X	X	X		X		X	X
Evaluating solutions and monitoring performance	X	X				X	X	X	X		X	X	X	X	X
Reaching correct solution	X				X					X			X		

Furthermore, ‘team learning’ was not covered by the reviewed assessments, which might be due to the role and purpose of the specific assessments in getting individual scores for students, rather than exploring the collaborative activity at the group level. In contrast, ‘exchanging information’ and ‘self-management and self-leadership’ skills were targeted by (almost) all assessments. This is not surprising, since most assessments used a mode of communication based on text messages, which might have made identifying and scoring skills such as ‘exchanging information’ easier. Similarly, the focus of the assessments on the individual contributions might have made scoring ‘self-management and self-leadership’ easier.

Table 5.3 also shows that in Leadership component, most assessments (n = 12) covered the skill ‘organising activities and resources’. It is worth noting that the assessments that covered all three skills in Leadership component (n = 4), facilitated collaboration between human participants (either through video or free-form text communication), and employed manual scoring of student interactions. It could be therefore hypothesised that skills within the Leadership component are elicited easier when using more authentic assessment environments. In Problem-solving component, most assessments (n = 14) covered the skill ‘brainstorming and identifying problems.’ In contrast, very few assessments (n = 4) covered the skill ‘reaching correct solution’. It is worth noting that only one assessment was found to cover all skills in Problem-solving component, however, no assessment was found to cover all skills detailed in the coding template.

Apart from the communication and scoring approaches that could facilitate or hinder the identification of students’ skills, another reason for only scarcely covering certain skills could be the fact that they are not being taught in the classroom. Finally, it might not be possible to develop one assessment, especially computer-simulated, that covers all facets of CPS competence. Therefore, other types of assessment data are needed to elicit skills that have not been covered by the computer-simulated, scenario-based assessment tasks reviewed.

5.5.3 What strategies for validating CPS competence measures are reported?

From the 26 articles in the review, 6 were excluded from this analysis, since they did not report the validation of an assessment, so this part uses only 20[20]. These acted as the units of analysis and were coded. Table 5.4 presents the types of validity evidence reported in the articles.

Table 5.4. Validity evidence reported in articles under review

	Validity aspects					
Author (Year)	1	2	3	4	5	6
Andrews and Forsyth (2020)	X			X	X	
Camacho-Morles et al. (2019)			X	X	X	
Care and Griffin (2014)	X	X				
Harding et al. (2017)	X		X	X		X
Harding and Griffin (2016)	X		X	X		
Herborn et al. (2020)			X		X	
Krkovic et al. (2016)			X	X		
Kuo et al. (2020)	X		X	X		
Lin et al. (2015)	X		X			
Nouri et al. (2017)	X				X	
Rosen (2015)	X				X	X
Rosen and Foltz (2014)	X		X		X	
Scoular and Care (2019)	X			X		
Scoular et al. (2020)	X			X		
Scoular and Care (2020)	X					
Siddiq and Scherer (2017)		X				
Stadler et al. (2019)				X	X	
Stadler et al. (2020b)				X	X	
Sun et al. (2020)	X		X		X	
Yuan et al. (2019)	X		X	X		

Notes: Articles are presented in alphabetical order using first author's name. Validity aspects 1-6: 1 = Content, 2 = Substantive, 3 = Structural, 4 = Generalisability, 5 = External, 6 = Consequential.

[20] The 20 articles represented 13 out of 15 assessments.

Table 5.4 shows that **content** validity evidence was the most frequently reported (n = 14). Articles evaluated the quality of items by reporting item fit statistics and item difficulties based on item response theory approaches (e.g., Harding et al., 2017; Scoular & Care, 2020). Others reported the frequency of codes reflecting each CPS skill targeted to ensure that there was an adequate number of observations to cover CPS competence concept (e.g., Andrews-Todd & Forsyth, 2020; Nouri et al., 2017). The second most widely evaluated validity aspect was **generalisability** (n = 11). Two reliability indices, item separation and person separation, were found to be reported within articles following item response theory approaches. Alternatively, values such as inter-rater reliability and Cronbach alpha values were reported.

Compared to content and generalisability validity evidence, **substantive** and **consequential** validity evidence was only very scarcely reported in the reviewed articles. This is not surprising, since researchers can use basic statistics to evaluate content and generalisability validity, but to evaluate evidence for substantive and consequential validity, it is argued that researchers have a lot of new types of data to collect and understand. For example, Siddiq and Scherer (2017) used "think-aloud" protocols to uncover the processes related to solving tasks collaboratively and examine the usability and authenticity of an assessment. Furthermore, social consequences resulting from test score use and interpretation are difficult to be evaluated. Harding et al. (2017) examined fairness between different test forms when students from different countries completed a CPS assessment and highlighted that further research is needed to ensure that students are not advantaged or disadvantaged by linguistic, communicative, cognitive, cultural, or other characteristics. Given that policy makers might attempt to include CPS competence assessment in the curriculum, especially after the recent publication of PISA 2015 CPS results, consequential validity evidence of CPS competence assessments and measures is important to be evaluated. Policy makers need to be aware of risks to prevent assessment starting to drive the curriculum rather than supporting it. For example, getting students trained to complete artificial exercises to score high marks for high-stakes assessment could build up unrealistic expectations of what collaboration is for students (Rosen, 2015).

Table 5.4 shows that half of the articles (n = 10) evaluated evidence about **structural** validity. Dimensionality of CPS competence construct was largely examined by means of item response theory or factor analysis approaches. For example, Kuo et al. (2020) found that multidimensional depictions (three- or four-dimensional) adequately described their CPS competence measures. Using a two-dimensional measurement model, consisting of a social/collaborative and a cognitive/problem-solving dimension, Harding et al. (2017) found evidence that CPS competence construct could be considered as multidimensional. Finally, almost half of the articles (n = 9) evaluated evidence about **external** validity. The external measures correlated with CPS competence measures were mainly self-reported constructs, such as perceived teamwork skills or motivation, from student questionnaires (e.g., Camacho-Morles et al., 2019; Rosen & Foltz, 2014; Stadler, Herborn, et al., 2020b; Sun et al., 2020). Other less frequently used external measures were students' performance in traditional subjects and students' teamwork skills as perceived by their teachers (e.g., Andrews-Todd & Forsyth, 2020; Stadler, Herborn, et al., 2020b).

Alternatively, a few articles have attempted to determine whether assessing students with the use of computer-simulated partners represented the way students would interact with human partners (e.g., Herborn et al., 2020; Rosen, 2015). However, the generalisability of their results is limited due to the constrained communication environments they utilised, which limited real-life unrestricted collaboration between humans. Consequently, the extent to which CPS competence captured using such constrained CPS tasks is consistent with, or different from, CPS competence exhibited in real human interactions is still to be determined.

5.6 Discussion

In this article, 15 assessments of students' CPS competence represented in 26 different articles were systematically reviewed. There are several results from this analysis, which have implications for educational policy, practice, and research.

5.6.1 Overview of collaborative problem-solving competence assessments and their characteristics

Assessments targeted almost exclusively secondary school students, pointing to the lack of tests measuring primary and higher education students' CPS competence, which future assessments could focus on. Results also showed a variation in the content of the problem-solving tasks. Harding et al. (2017) found that students scored differently when tasks were framed in the mathematics context compared to content-free context, and it was argued that new CPS assessments should be designed to take the relationship between mathematics and CPS competence into account. Therefore, understanding how CPS competence relates to other content areas is needed to improve CPS assessments.

Furthermore, placing students in pairs for the assessment of CPS competence was preferred over relatively more complex groups, such as triads or larger groups. Group size is suggested to influence the exhibition of members' CPS skills in the group (Hao et al., 2017; Salas et al., 2017). One potential advantage of dyads is that they provide more outcome data points than other arrangements (Fiore et al., 2017). Increasing the group size may also lead to increased social loafing phenomenon, that is, a group member being less motivated and relying on others to contribute to the team outcome (Fiore et al., 2017). Since it is not clear how different group sizes affect collaboration outcomes, future research could compare different group arrangements.

The approach of replacing real humans with computer-simulated partners in the groups has been previously criticised for lacking authenticity and deviating from ecologically valid CPS activities (Graesser et al., 2018; Rosen, 2015; Scoular et al., 2017; Siddiq & Scherer, 2017). Specifically, the way students interact with computer-simulated partners was argued to be limited in matching how students interact with real students (Scoular & Care, 2020). One reason for that is because the dynamics of human interaction cannot be perfectly captured with computer-simulated partners (Rosen, 2015). In one assessment, for example, it was highlighted that it was difficult to design response formats for computer-simulated partners that were both logical and "human-like" (Lin et al., 2015). In addition, human collaborators can propose unusual or exceptional solutions, which cannot be included in a system employing computer-simulated partners (Rosen, 2015). One question that therefore arises is whether a CPS competence construct measured in this way can represent competence that transfers to real-life CPS.

Although the phrase-chat communication approach was claimed to provide a manageable way to track communication allowing for automated scoring (Herborn et al., 2020; Rosen, 2015), it raised several validity-related concerns (Andrews-Todd & Forsyth, 2020; Scoular, Eleftheriadou, et al., 2020; Scoular & Care, 2020; Sun et al., 2020). Specifically, it was argued that phrase-chat could not perfectly capture the full range of student interactions (e.g., unexpected responses, negotiation), since the richness of dialogue that could occur in open collaborative environments was lost (e.g., Andrews-Todd & Forsyth, 2020; Scoular, Eleftheriadou, et al., 2020; Scoular & Care, 2020). The difficulty in capturing negotiations has also been highlighted, as it often takes a multi-turn exchange between group members for negotiation to happen, which was not permitted when tasks used phrase-chat communication (Graesser et al., 2018). Very few assessments were found to facilitate more authentic interaction among group members, which points to the fact that their potential has not been fully exploited yet. Consequently, future developments in task design may target facilitating more authentic CPS competence assessments.

Computer-simulated, scenario-based tasks can capture an abundance of data including students' actions and messages. However, the content of messages has been so far largely ignored when scoring student interaction using algorithms. When scoring is focused almost exclusively on the frequency of occurrences of actions or chats (i.e., counts of messages) without analysing chat content itself, an incomplete picture of the nature of what is being measured is provided. As Nouri et al. (2017) found, more frequent communication did not guarantee success in solving the problem, since some students based their interactions on trial-and-error strategies. Consequently, more substantial analysis of the communication content is needed.

5.6.2 Facets of collaborative problem-solving competence

By revealing the gaps in CPS competence targeting by existing assessments, this article contributes to a continued debate on developing assessments that will result in measures more authentically well aligned with the whole concept. It is important to note that the reviewed assessments followed common practices in educational measurement such as the use of a construct-models that break down a concept into minute skills. This work can be criticised, and its limitations can be exposed, because of this restriction.

Several skills related to Teamwork and Communication components, such as 'active listening', although still considered as important in CPS competence by many, were scarcely or not at all covered within the reviewed assessments. It is therefore necessary to understand what might hinder capturing these skills and further explore how to potentially measure them using alternative task design features. In addition, the mode of communication, whether students are collaborating with other human participants, how authentic the assessment environment is, the scoring approaches, or whether students have previously been taught the targeted skills in the classroom, may influence what is targeted for assessment.

For example, in multiple-choice communication formats with the use of computer-simulated partners, it has been argued that there is less opportunity to capture the wide range of behaviours implicit in CPS competence due to constraints on choice and sequence of actions and on the chat (Scoular et al., 2017). There are also no opportunities for lengthy conversation threads to handle negotiation and building on each other's ideas (Andrews-Todd & Forsyth, 2020; Graesser et al., 2018). In contrast, individualised computer-simulated assessments might have facilitated the identification of skills such as 'exchanging information', since they could be more easily captured in automatic ways.

Task design features can affect what is made available for measurement and what is ultimately measured (Nouri et al., 2017). For example, restricting the communication or deciding not to interpret the communication content are some task design decisions that can make analyses easier compared to an open environment (Andrews-Todd & Kerr, 2019). However, such approaches may introduce issues related to oversimplification (Webb & Gibson, 2015) by capturing only straightforward or relatively easy to measure indicators. To get more comprehensive and authentic measures of students' CPS competence, what is needed from future research is a more holistic understanding of why students respond in the way they do in the collaborative assessments and what aspects need to be scored.

Given the complexity of CPS competence concept, prioritising some (possibly small-scale) qualitative studies, where students explain their response processes, could help get that in-depth analysis. For instance, observations in authentic situations and think-aloud protocols can help researchers during instrument development with scoring aspects that have not been captured by the assessments to date (Willis, 2005).

5.6.3 Validity evidence

The results also revealed that evidence for the external, substantive, and consequential validity aspects is only scarcely examined. This is problematic as it is difficult to evaluate CPS competence measures when such evidence is not provided.

Current attempts to validate CPS competence measures have focused mainly on evidence from psychometric models reporting item fit statistics, dimensionality tests, and reliability indices. The lack of validation using authentic problem-solving tasks utilised by teachers in classrooms is still apparent. Results showed that most external validity evidence is based on correlating student scores with self-assessments. This approach is likely to suffer from weak external validity in the sense that a student can score well on an assessment without being a competent collaborative problem solver in an authentic situation. Evidence for external validity will need to go beyond the measurement itself and even beyond other academic measures, since students who are good enough to do mathematics or science, might be also good enough to answer CPS assessments. Additionally, in the case of very bright and intuitive students, there is a chance of “gaming” the task if students respond based on guessing what the desired responses or outcomes can be, rather than what they would do under regular conditions (Oliveri et al., 2017).

While there is an element of external validity in correlating student scores with other perceived attitudes and outcomes, what is missing from current literature is observing students in real situations. For CPS competence measures to be valid, it would be essential for constructs assessed though computer-simulated, scenario-based tasks to be identical to those assessed in real-life problem-solving situations (Webb & Gibson, 2015). It is important to deepen our understanding of whether the CPS assessments reflect students’ competence in CPS, or whether students’ intelligence, intuition, or even test-savviness, are sufficient for doing well in these assessments.

For instance, are students who score well on computer-simulated, scenario-based tasks, also competent collaborative problem-solvers in a real, authentic situation? To answer that, researchers are encouraged to conduct (possibly small-scale) qualitative studies using authentic problem-solving tasks as external measures/assessments. This is not an easy task and requires a lot of resources, however, there are examples of authentic assessments of actual practical problem solving in the literature that could be used in collaborative situations.

Social consequences of test score use, and interpretation, were also found to be inadequately addressed. When building assessment instruments, it is important to consider that they might get out of control and start to drive what is happening instead of measuring it. For example, concerns have been raised about the possibility of countries adopting a computer-simulated approach in high-stakes assessment of CPS competence in schools (Webb & Gibson, 2015).

Education systems in different parts of the world have recently shown interest in curriculum reform and have progressively begun to incorporate CPS into their curricula and teaching pedagogies (Care, Anderson, et al., 2016). What needs to be carefully considered are the social consequences of introducing CPS assessments as high-stakes assessments or as part of the curriculum in schools in general. In this case, rather than teaching students to solve problems in the classroom, which is more authentic, there is a risk that teachers might teach students to the test, encouraging them to practice their CPS skills by doing artificial exercises in computer-simulated environments. If educators rely exclusively on using assessments with computer-simulated partners, there is a risk that students might build up expectations about interactions that deviate from natural human communication, given that computer-simulated partners cannot perfectly capture the dynamics of interaction between human students (Rosen, 2015).

Bias could also be introduced if students realise that they are working with computer-simulated partners, and change their behaviour (Krkovic et al., 2016). Additionally, assessment environments that deviate from natural human communication may cause distraction or even irritation to students (Rosen, 2015). For that reason, it has been suggested to avoid using these tasks in isolation. Instead, a blended or tiered approach to assessment have been recommended. In the former, the simulated, scenario-based tasks are used in combination with other forms of assessments (Oliveri et al., 2017), while in the latter, human-to-human collaboration is considered the optimal approach and human-to-computer-simulated partner is an optional approach used to at least capture parts of CPS competence, where the optimal approach is not possible (Scoular et al., 2017). Each approach to CPS assessment could be effective for different educational purposes, which is likely to be influenced by the nature and stakes of assessment.

Furthermore, CPS competence is rarely explicitly included as part of school curricula, which raises concerns about the validity of assessment inferences, since such assessments can introduce differences in opportunity to learn (Ercikan & Oliveri, 2016). Specifically, the derived CPS competence measures reflect whether students have had the opportunity to develop CPS competence outside of the schooling contexts or other factors that may not be the focus of the assessment (e.g., general intelligence), and consequently, limited connections can be made between schooling and outcomes on CPS competence assessments (Ercikan & Oliveri, 2016).

Overall, the social consequences of the use of CPS assessments need to be carefully considered. As argued, there are risks involved when students practise interacting with artificial exercises, rather than in real life scenarios. Policy makers need to be aware of those risks to prevent assessment starting to drive the curriculum rather than supporting it. Finally, future research needs to focus on evaluating evidence for the validity aspects that have not been fully covered in the literature yet.

5.6.4 Limitations

There are several limitations that need to be addressed. First, the systematic literature review methodology included only sources published in English and those in peer-reviewed journals. These choices influenced the findings and may have omitted quality work in other languages and work not published in academic journals. A further limitation is the inclusion of empirical articles only, which was necessary for answering the third research question (What strategies for validating CPS competence measures are reported?). Future research studies could offer a more extended analysis of the remaining literature. Nevertheless, the comprehensive overview presented in this article can serve as a starting point for future reviews and for those who are new to the assessment of CPS competence.

5.7 Conclusion

Assessments aimed to measure students' CPS competence has been increasing in the educational literature. Since existing literature reviews were found to be limited due to their focus, this article offers a state-of-the-art in the assessment of CPS competence and can form the basis for future research. In this article, assessments of students' CPS competence using computer-simulated, scenario-based tasks were systematically reviewed. Then, a critical evaluation of the CPS competence-components they targeted was reported along with the validity evidence they provided. A systematic review methodology was followed to address transparency and replicability (Gough et al., 2016; Petticrew & Roberts, 2006). In conclusion, this article shed light into some of the limitations in current assessments of CPS competence and highlighted gaps in the validity evidence concerning CPS competence measures. In light of this, it is recommended that educational policy should be sensitive to the authenticity of assessments and the social consequences of test score use and interpretation. It may be more informative and productive to research students in real-life situations to inform assessment development and subsequently get measures more authentically well aligned with CPS in authentic situations.

5.8 References

Articles (n = 26) included in the review are marked with an asterisk (*).

Alexander, P. A. (2020). Methodological Guidance Paper: The Art and Science of Quality Systematic Reviews. *Review of Educational Research*, *90*(1), 6–23. https://doi.org/10.3102/0034654319854352

*Andrews-Todd, J., & Forsyth, C. M. (2020). Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Computers in Human Behavior*, *104*, 105759. https://doi.org/10.1016/j.chb.2018.10.025

*Andrews-Todd, J., Jackson, G. T., & Kurzum, C. (2019). Collaborative Problem Solving Assessment in an Online Mathematics Task. *ETS Research Report Series*, *2019*(1), 1–7. https://doi.org/10.1002/ets2.12260

Andrews-Todd, J., & Kerr, D. (2019). Application of Ontologies for Assessing Collaborative Problem Solving Skills. *International Journal of Testing*, *19*(2), 172–187. https://doi.org/10.1080/15305058.2019.1573823

*Camacho-Morles, J., Slemp, G. R., Oades, L. G., Morrish, L., & Scoular, C. (2019). The role of achievement emotions in the collaborative problem-solving performance of adolescents. *Learning and Individual Differences*, *70*, 169–181. https://doi.org/10.1016/j.lindif.2019.02.005

Camara, W., O'Connor, R., Mattern, K., & Hanson, M. A. (2015). *Beyond Academics: A Holistic Framework for Enhancing Education and Workplace Success* (Act Research Report Series). ACT Inc.

Care, E., Anderson, K., & Kim, H. (2016). *Visualizing the Breadth of Skills Movement Across Education Systems*. Center for Universal Education at the Brookings Institution.

*Care, E., & Griffin, P. (2014). An Approach to Assessment of Collaborative Problem Solving. *Research and Practice in Technology Enhanced Learning*, *9*(3), 367–388.

Cukurova, M., Luckin, R., Millán, E., & Mavrikis, M. (2018). The NISPI framework: Analysing collaborative problem-solving from students' physical interactions. *Computers & Education*, *116*, 93–109. https://doi.org/10.1016/j.compedu.2017.08.007

*De Boeck, P., & Scalise, K. (2019). Collaborative Problem Solving: Processing Actions, Time, and Performance. *Frontiers in Psychology*, *10*, 1280. https://doi.org/10.3389/fpsyg.2019.01280

Ercikan, K., & Oliveri, M. E. (2016). In Search of Validity Evidence in Support of the Interpretation and Use of Assessments of Complex Constructs: Discussion of Research on Assessing 21st Century Skills. *Applied Measurement in*

Education, *29*(4), 310–318. https://doi.org/10.1080/08957347.2016.1209210

Fiore, S. M., Graesser, A., Greiff, S., Griffin, P., Gong, B., Kyllonen, P., Massey, C., O'Neil, H., Pellegrino, J., Rothman, R., Soulé, H., & von Davier, A. (2017). *Collaborative Problem Solving: Considerations for the National Assessment of Educational Progress*. National Center for Education Statistics. http://orbilu.uni.lu/handle/10993/31897

Gough, D., Oliver, S., & Thomas, J. (2016). *An introduction to systematic reviews* (2nd ed.). SAGE Publications Ltd.

Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the Science of Collaborative Problem Solving. *Psychological Science in the Public Interest*, *19*(2), 59–92. https://doi.org/10.1177/1529100618808244

Graesser, A. C., Greiff, S., Stadler, M., & Shubeck, K. T. (2020). Collaboration in the 21st century: The theory, assessment, and teaching of collaborative problem solving. *Computers in Human Behavior*, *104*, 106134. https://doi.org/10.1016/j.chb.2019.09.010

Greiff, S., & Kyllonen, P. (2016). Contemporary Assessment Challenges: The Measurement of 21st Century Skills. *Applied Measurement in Education*, *29*(4), 243–244. https://doi.org/10.1080/08957347.2016.1209209

Griffin, P., & Care, E. (2014). *Assessment and teaching of 21st century skills: Methods and approach*. Springer.

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

Hao, J., Liu, L., von Davier, A., & Kyllonen, P. (2017). Initial Steps Towards a Standardized Assessment for Collaborative Problem Solving (CPS): Practical Challenges and Strategies. In A. von Davier, M. Zhu, & P. Kyllonen (Eds.), *Innovative Assessment of Collaboration* (pp. 135–156). Springer International Publishing. https://doi.org/10.1007/978-3-319-33261-1_9

*Harding, S.-M. E., & Griffin, P. (2016). Rasch Measurement of Collaborative Problem Solving in an Online Environment. *Journal of Applied Measurement*, *1*(17), 35–53.

*Harding, S.-M. E., Griffin, P., Awwal, N., Alom, B. M., & Scoular, C. (2017). Measuring Collaborative Problem Solving Using Mathematics-Based Tasks. *AERA Open*, *3*(3), 1–19. https://doi.org/10.1177/2332858417728046

*Herborn, K., Mustafić, M., & Greiff, S. (2017). Mapping an Experiment-Based Assessment of Collaborative Behavior Onto Collaborative Problem Solving in PISA 2015: A Cluster Analysis Approach for Collaborator Profiles. *Journal of Educational Measurement*, *54*(1), 103–122. https://doi.org/10.1111/jedm.12135

*Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? *Computers in Human Behavior*, *104*, 105624. https://doi.org/10.1016/j.chb.2018.07.035

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A Framework for Teachable Collaborative Problem Solving Skills. In P. Griffin & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 37–56). Springer Netherlands. https://doi.org/10.1007/978-94-017-9395-7_2

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

*Krkovic, K., Wüstenberg, S., & Greiff, S. (2016). Assessing Collaborative Behavior in Students: An Experiment-Based Assessment Approach. *European Journal of Psychological Assessment*, *32*(1), 52–60. https://doi.org/10.1027/1015-5759/a000329

*Kuo, B.-C., Liao, C.-H., Pai, K.-C., Shih, S.-C., Li, C.-H., & Mok, M. M. C. (2020). Computer-based collaborative problem-solving assessment in Taiwan. *Educational Psychology*, *40*(9), 1164–1185. https://doi.org/10.1080/01443410.2018.1549317

*Li, C.-H., & Liu, Z.-Y. (2017). Collaborative Problem-Solving Behavior of 15-Year-Old Taiwanese Students in Science Education. *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(10), 6677–6695. https://doi.org/10.12973/ejmste/78189

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

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus Article: On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research & Perspective*, *1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Annals of Internal Medicine*, *151*(4), 264–269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

*Nouri, J., Åkerfeldt, A., Fors, U., & Selander, S. (2017). Assessing Collaborative Problem Solving Skills in Technology-Enhanced Learning Environments – The PISA Framework and Modes of Communication. *International Journal of Emerging Technologies in Learning (IJET)*, *12*(04), 163. https://doi.org/10.3991/ijet.v12i04.6737

OECD. (2017a). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. OECD Publishing. https://doi.org/10.1787/9789264281820-en

OECD. (2017b). *PISA 2015 Results (Volume V): Collaborative problem solving*. OECD Publishing. https://doi.org/10.1787/9789264285521-en

Oliveri, M. E., Lawless, R., & Molloy, H. (2017). A Literature Review on Collaborative Problem Solving for College and Workforce Readiness. *ETS Research Report Series*, *2017*(1), 1–27. https://doi.org/10.1002/ets2.12133

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell.

*Polyak, S. T., von Davier, A., & Peterschmidt, K. (2017). Computational Psychometrics for the Measurement of Collaborative Problem Solving Skills. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.02029

*Rosen, Y. (2014). Comparability of Conflict Opportunities in Human-to-Human and Human-to-Agent Online Collaborative Problem Solving. *Technology, Knowledge and Learning*, *19*, 147–164. https://doi.org/10.1007/s10758-014-9229-1

*Rosen, Y. (2015). Computer-based Assessment of Collaborative Problem Solving: Exploring the Feasibility of Human-to-Agent Approach. *International Journal of Artificial Intelligence in Education*, *25*(3), 380–406. https://doi.org/10.1007/s40593-015-0042-3

*Rosen, Y., & Foltz, P. (2014). Assessing collaborative problem solving through automated technologies. *Research and Practice in Technology Enhanced Learning*, *9*, 389–410.

Salas, E., Reyes, D., & Woods, A. (2017). The Assessment of Team Performance: Observations and Needs. In A. von Davier, M. Zhu, & P. Kyllonen (Eds.), *Innovative Assessment of Collaboration* (pp. 21–36). Springer International Publishing. https://doi.org/10.1007/978-3-319-33261-1_2

*Scoular, C., & Care, E. (2019). A Generalized Scoring Process to Measure Collaborative Problem Solving in Online Environments. *Educational Assessment*, *24*(3), 213–234. https://doi.org/10.1080/10627197.2019.1615372

*Scoular, C., & Care, E. (2020). Monitoring patterns of social and cognitive student behaviors in online collaborative problem solving assessments. *Computers in Human Behavior*, 105874. https://doi.org/10.1016/j.chb.2019.01.007

Scoular, C., Care, E., & Hesse, F. W. (2017). Designs for Operationalizing Collaborative Problem Solving for Automated Assessment. *Journal of Educational Measurement*, *54*(1), 12–35. https://doi.org/10.1111/jedm.12130

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

*Scoular, C., Eleftheriadou, S., Ramalingam, D., & Cloney, D. (2020). Comparative analysis of student performance in collaborative problem solving: What does it tell us? *Australian Journal of Education*. https://doi.org/10.1177/0004944120957390

*Siddiq, F., & Scherer, R. (2017). Revealing the processes of students' interaction with a novel collaborative problem solving task: An in-depth analysis of think-aloud protocols. *Computers in Human Behavior*, *76*, 509–525. https://doi.org/10.1016/j.chb.2017.08.007

*Stadler, M., Herborn, K., Mustafić, M., & Greiff, S. (2019). Computer-Based Collaborative Problem Solving in PISA 2015 and the Role of Personality. *Journal of Intelligence*, *7*(3), 15. https://doi.org/10.3390/jintelligence7030015

*Stadler, M., Herborn, K., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: An investigation of the validity of the PISA 2015 CPS tasks. *Computers & Education*, *157*, 103964. https://doi.org/10.1016/j.compedu.2020.103964

*Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education*, *143*, 103672. https://doi.org/10.1016/j.compedu.2019.103672

Thomas, J., Brunton, J., & Graziosi, S. (2010). *EPPI-Reviewer 4: Software for research synthesis*. Social Science Research Unit, UCL Institute of Education.

Vista, A., Care, E., & Awwal, N. (2017). Visualising and examining sequential actions as behavioural paths that can be interpreted as markers of complex behaviours. *Computers in Human Behavior*, *76*, 656–671. https://doi.org/10.1016/j.chb.2017.01.027

von Davier, A. (2017). Computational Psychometrics in Support of Collaborative Educational Assessments: Computational Psychometrics. *Journal of Educational Measurement*, *54*(1), 3–11. https://doi.org/10.1111/jedm.12129

von Davier, A., Zhu, M., & Kyllonen, P. (Eds.). (2017). *Innovative Assessment of Collaboration*. Springer International Publishing. https://www.springer.com/gb/book/9783319332598

Webb, M., & Gibson, D. (2015). Technology enhanced assessment in complex collaborative settings. *Education and Information Technologies*, *20*(4), 675–695. https://doi.org/10.1007/s10639-015-9413-5

Willis, G. B. (2005). *Cognitive interviewing a tool for improving questionnaire design*. SAGE.

Wolfe, E. W., & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part I - instrument development tools. *Journal of Applied Measurement*, *8*(1), 97–123.

*Yuan, J., Xiao, Y., & Liu, H. (2019). Assessment of Collaborative Problem Solving Based on Process Stream Data: A New Paradigm for Extracting Indicators and Modeling Dyad Data. *Frontiers in Psychology*, *10*, 369. https://doi.org/10.3389/fpsyg.2019.00369

Chapter 6 Dimensionality and validity of the PISA 2015 collaborative problem-solving competence construct

6.1 Abstract

The topic of collaborative problem solving (CPS) competence has received attention in educational literature, especially after the release of the Programme for International Student Assessment (PISA) 2015 results, where 15-year-olds'competence to work in collaborative settings has been assessed across countries. Numerous definitions and frameworks have been developed to assess students' CPS competence, mainly at the individual/student level and with the use of computer-assisted technologies. These are generally multidimensional, distinguishing among sets of skills (e.g., shared understanding, taking actions, team organisation). However, several computer-based assessments constrained the communication among group members, raising several questions regarding validity, especially external and consequential validity aspects. In this paper, the validity of the PISA CPS competence measure was evaluated using the Rasch measurement framework. The multidimensional character of CPS competence was first investigated using students' scored responses to the PISA 2015 CPS assessment. Measurement invariance, and the association of the derived CPS competence measures with related collaboration constructs, were also tested. Results showed that the CPS items collectively measure a unidimensional model of CPS competence, while the associations with other theoretically relevant variables did not support external validity. Such results call for more caution in interpreting results based on PISA CPS competence measure to inform policy and practice. Results also strengthen the case that further evidence is needed for the external aspect of validity to become more prominent in future research studies. Further reflections about scale modifications, policy and practice implications and suggestions for future research are discussed.

Keywords: collaborative problem solving, assessment, measurement, validity, Rasch model, PISA

6.2 Introduction

Collaborative problem solving (CPS) competence has become increasingly researched and reported in the educational literature across many domains such as mathematics as well as across disciplines (Care et al., 2018). The need to prepare students for careers that require them to work effectively in groups and to apply problem-solving skills in social situations has driven interest in teaching and assessing students' CPS competence (Griffin et al., 2012; National Research Council, 2011; OECD, 2017a). In the education sector, the recent inclusion of an assessment of CPS competence in the Programme for International Student Assessment (PISA) study in 2015 highlights the interest in this concept internationally. In PISA's conceptualisation, three main CPS competencies were identified: Establishing and maintaining shared understanding, Taking appropriate action to solve the problem, and Establishing and maintaining team organisation (OECD, 2017a). Despite its multidimensional nature, a few studies have so far investigated the validity of CPS competence measures in general and dimensionality in particular (Gao et al., 2022; Harding et al., 2017; Harding & Griffin, 2016; Kuo et al., 2020; Yuan et al., 2019).

Additionally, for the purposes of assessing students' CPS competence, constrained computer-based environments have been increasingly used, as with PISA 2015, to provide standardised environments that facilitate capturing and scoring students' social interactions. In a recent systematic literature review (CPS measurement review; Chapter 5), 15 different assessments of students' CPS competence (including PISA 2015 CPS assessment) were reviewed to reveal which aspects of CPS competence were measured. Assessments adopting test design features such as the replacement of human group members with computer-simulated partners have been criticised for deviating from naturalistic, ecologically valid CPS activities (Graesser et al., 2018; Scoular et al., 2017), raising, in turn, concerns about external and consequential validity. Only a limited number of studies have so far examined evidence to shed light on these validity aspects (Herborn et al., 2020; Nouri et al., 2017; Rosen, 2015; Rosen & Foltz, 2014; Stadler, Herborn, et al., 2020a). My CPS measurement review (Chapter 5) found no articles using the PISA 2015 data for CPS to specifically explore external and consequential validity issues.

So far, only a limited number of studies have empirically investigated the multidimensional nature of CPS competence (Harding et al., 2017; Harding & Griffin, 2016; Herborn et al., 2020; Krkovic et al., 2016; Kuo et al., 2020; Sun et al., 2020; Yuan et al., 2019). For instance, Harding et al. (2017) found evidence supporting a two-dimensional measurement model with a social and a cognitive dimension. In another study, Krkovic et al. (2016) confirmed that their theoretically hypothesised five-factor model, consisting of two problem-solving dimensions (knowledge acquisition and knowledge application) and three collaboration dimensions (questioning, asserting, and requesting), was statistically adequate. A noticeable gap can be seen in using secondary PISA 2015 data to explore the multidimensionality of CPS competence and the current paper contributes towards that end.

Given that the publishing of the PISA 2015 CPS results will likely increase the attention received from researchers, educators, and policy makers on students' CPS competence, this, in turn, might drive policy decisions and curricula revisions. Evidence on the adequacy and appropriateness of interpretations and actions based on PISA CPS competence scores is therefore needed. This paper investigates validity aspects of CPS competence measure(s) constructed using students' responses to the PISA 2015 CPS assessment. Based on Messick's (1989, 1995) unified validity definition, validation is a continuing process referring to the accumulation of evidence to support validity arguments. More specifically, the current paper adds value to the existing literature and contributes to knowledge about the validity of PISA's CPS competence measure via three aims: (i) to explore whether there is evidence of the concept's multidimensionality in students' responses to the PISA 2015 CPS assessment, (ii) to explore measurement invariance by gender, and (iii) to explore the association of the derived CPS competence measure(s) with relevant collaboration constructs.

The paper proceeds as follows. The next section presents the background of the paper including the research questions, followed by an overview of methods. Results are then reported, followed by a discussion and the conclusion.

6.3 Background

6.3.1 Defining collaborative problem-solving competence

Problem solving has been assessed for several decades, following Polya's (1945) well-known problem-solving process which consisted of four steps: understanding the problem, devising a plan, carrying out the plan, and looking back. Driven by the perceived needs of policy, PISA 2015 study aimed to measure, and consequently ensure that students are equipped with, skills to meet the CPS demands of their future careers. For the purposes of PISA 2015 CPS assessment, CPS competency was defined as "the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution" (OECD, 2017a, p. 134).

Apart from the PISA 2015 CPS framework, several other CPS frameworks have been developed by researchers and organisations defining the concept and its main components (e.g., Andrews-Todd & Forsyth, 2020; Hesse et al., 2015; Oliveri et al., 2017; O'Neil et al., 2003; Sun et al., 2020). Different authors conceived a range of different skills (or processes) associated with CPS competence, but in general they shared some commonalities in their conceptualisations. One of those is the multidimensional character of CPS competence, which consisted of a suite of skills (or processes) needed for effective teamwork in service of problem solving, and which are organised in various dimensions. The structure ranged from two dimensions, a cognitive dimension associated with problem-solving skills and a social dimension associated with collaboration skills (Andrews-Todd & Forsyth, 2020; Hesse et al., 2015; O'Neil et al., 2003) to three (Sun et al., 2020) and four dimensions (Oliveri et al., 2017). Most importantly, the analysis of CPS competence into its contributing skills (or processes) was deemed necessary for facilitating the design of assessment tasks and item development.

The various descriptions of CPS competence provide an opportunity to interrogate the nature of the concept and the degree to which these various frameworks, and subsequently the assessments, cover the same skills (Scoular & Care, 2020). Most of the assessments in the systematic review (Chapter 5; Research paper 2) were found to measure a limited spectrum of CPS skills, when mapped on existing CPS frameworks to investigate construct representation and targeting. Specifically, skills such as active listening, audience awareness, team learning, and team empowerment, were only scarcely or not at all covered. In another study, Scoular, Eleftheriadou et al. (2020) found that skills related to negotiation and audience awareness were not well represented across three assessments of students' CPS competence (including PISA 2015). Such findings indicate that test developers have had limited success in eliciting these skills in computer-based assessment contexts to date.

6.3.2 The PISA 2015 collaborative problem-solving framework

The definition of CPS competence, for the purposes of the PISA 2015 CPS assessment, incorporates three core competencies unique to PISA's CPS: Establishing and maintaining shared understanding, Taking appropriate action to solve the problem, and Establishing and maintaining team organisation (OECD, 2017a). These three newly conceptualised competencies are crossed with the four individual problem-solving processes (i.e., exploring and understanding, representing and formulating, planning and executing, and monitoring and reflecting) to create a matrix of 12 cells, each representing a specific CPS skill (Table 6.1). The individual problem-solving processes have been previously defined in the PISA 2012 framework following Polya's (1945) work related to problem solving in the context of mathematics.

Each item included in the CPS assessment is classified as targeting one of the CPS skills, and thus it can be mapped back to one of the three core competencies. Several points should be made in relation to construct representation and item distribution. Table 6.1 shows that items are not equally distributed across the three CPS competencies, e.g., about half of the items represent "Establishing and maintaining shared understanding”. OECD (2017a) justify the greater weight because “Establishing and maintaining shared understanding” and “Establishing and maintaining team organisation” focus specifically on collaborative skills, while “Taking appropriate action to solve the problem” focuses more on problem-solving behaviour within a collaborative context. Table 6.1 also shows that the CPS skill “Understanding roles to solve the problem” has not been targeted by any item in the assessment, raising concerns about the validity of the derived CPS competence measure (Messick, 1995). Other CPS skills (e.g., A2 and D2 in Table 6.1) have been targeted by a small number of items (n = 2 and 3, respectively), raising similar validity concerns. Therefore, the issue of construct underrepresentation, a situation in which essential aspects of a given construct have not been captured (American Educational Research Association et al., 2014), needs to be further investigated.

The chosen scoring method for PISA 2015 CPS assessment assigns a single overall score to each item response, which contributes to a student’s overall CPS competence measure. Similarly, to interpret student performance on CPS competence as a single construct, researchers have used unidimensional models such as the Rasch model (e.g., Scoular, Eleftheriadou, et al., 2020; Scoular & Care, 2020). However, it has been argued that the richness and complexity of the CPS construct could possibly mean that there are several sub-dimensions that contribute to it (Scoular & Care, 2020). So far, the role of the overarching dimensional structure characterising CPS competence has been left unexplored. This paper aims to contribute to current knowledge by investigating whether there is evidence suggesting that the CPS competence measure, as defined and assessed by the PISA 2015 CPS assessment, should be multi-dimensional.

Table 6.1. PISA 2015 Collaborative problem-solving framework (OECD, 2017a)

	(1) Establishing and maintaining shared understanding	(2) Taking appropriate action to solve the problem	(3) Establishing and maintaining team organisation
(A) Exploring and understanding	(A1) Discovering perspectives and abilities of team members **20 items**	(A2) Discovering the type of collaborative interaction to solve the problem, along with goals **2 items**	(A3) Understanding roles to solve the problem **0 items**
(B) Representing and formulating	(B1) Building a shared representation and negotiating the meaning of the problem (common ground) **24 items**	(B2) Identifying and describing tasks to be completed **5 items**	(B3) Describing roles and team organisation (communication protocol/rules of engagement) **8 items**
(C) Planning and Executing	(C1) Communicating with team members about the actions to be/being performed **5 items**	(C2) Enacting plans **16 items**	(C3) Following rules of engagement (e.g., prompting other team members to perform their tasks) **14 items**
(D) Monitoring and reflecting	(D1) Monitoring and repairing the shared understanding **12 items**	(D2) Monitoring results of actions and evaluating success in solving the problem **3 items**	(D3) Monitoring, providing feedback and adapting the team organisation and roles **8 items**

Notes: Number of items in bold are own adaptation of the framework based on information found in PISA’s technical report (OECD, 2017c)

6.3.3 Assessment of collaborative problem-solving competence

Following the development of the CPS frameworks, the focus of the current literature is on the assessment of CPS competence. Features of existing computer-simulated, scenario-based assessments of CPS competence were critically appraised in the CPS measurement review (Chapter 5). Despite its novelty and theoretical considerations, I argued there that PISA's approach towards CPS assessment is limited in that it: (i) considers CPS only from the perspective of individuals' competence, (ii) requires students to interact with computer-simulated partners programmed to respond in certain ways, and (iii) constrains the exchange of communication messages by limiting the number of possible responses. Overall, the assessment has been critiqued for deviating from naturalistic, ecologically valid CPS activities (Cukurova, Luckin, Millán, et al., 2018; Graesser et al., 2018; Scoular et al., 2017). Such criticisms raise, in turn, concerns about the interpretation and consequences of test score use (Messick, 1995).

It has been argued that assessment settings utilising computer-simulated partners cannot perfectly capture the dynamics of interaction between two humans, since collaboration in those settings deviates from natural human communication and computer-simulated partners cannot adjust to idiosyncratic characteristics of humans (Rosen & Foltz, 2014). For example, human partners can propose unusual, exceptional solutions during collaboration, but such a process cannot be included in a system following an algorithm (Rosen & Foltz, 2014). Researchers trying to apply the PISA CPS assessment approach in their studies also highlighted the difficulty to design a problem scenario assessing "Establishing and maintaining team organisation", mostly due to the difficulty to design a response format that was sufficiently 'human-like' for the computer-simulated partner (Lin et al., 2015). Given that communication is identified as having a central role in CPS competence, it is considered unlikely that assessment scenarios utilising computer-simulated partners would be able to elicit sufficient behaviours for good representation of CPS competence (Scoular et al., 2017). A question that, therefore, arises from the above is whether the scores derived from such interactions are valid for measuring CPS competence.

In addition, for the purposes of the PISA 2015 CPS assessment, the communication among group members was constrained to a selection of predefined messages (OECD, 2017a). Assessment approaches that allow only predefined chat communication between group members have been also criticised for creating unrealistic collaboration environments and limiting natural collaboration conversations (Scoular et al., 2017; Scoular & Care, 2020). In addition, the lack of structure that characterises discourse between humans in comparison to the structure that the communication with a computer-simulated partner imposes on conversation, pose a major challenge in capturing meaningful utterances linked to CPS competence (Graesser et al., 2017; Scoular & Care, 2020). For instance, students in the PISA 2015 CPS assessment had no opportunities for lengthy conversations, to negotiate or build on each other's ideas, while personality and emotions of team members were also out of scope (Graesser et al., 2018). As students were not able to introduce new ideas other than the predefined ones, it could be further argued that spontaneity and creativity have also been constrained, and there was, therefore, less opportunity for them to be captured.

A few studies have so far attempted to investigate whether computer-simulated partners can validly replace humans as collaborative partners in CPS assessments (Herborn et al., 2020; Rosen, 2015). Although valuable, the generalisability of their results is limited, since significant constraints on the collaboration between human partners were still posed (e.g., no free-flowing conversation or face-to-face interaction). For CPS competence measures to be valid, it would be essential for constructs assessed though computer-simulated, scenario-based tasks to be identical to those assessed in real life problem-solving (Webb & Gibson, 2015). As shown in the CPS measurement review (Chapter 5), current attempts to validate CPS competence measures have focused mainly on evidence from psychometric analysis targeting the generalisability aspect of construct validity. To date, no studies were found to compare CPS competence measures derived from computer-based assessments with authentic collaborative situations. Controlling the characteristics of the group members and constraining the communication, were deemed necessary for the purposes of a large-scale international assessment such as PISA. Such controlled assessment tools, however, raise concerns regarding aspects of validity (particularly external and consequential), which is the most fundamental consideration in developing tests (American Educational Research Association et al., 2014). This paper contributes to current knowledge about the validity of PISA's CPS competence measure by investigating the extent to which this is associated with related constructs such as students' collaboration attitudes.

6.3.4 Research questions

There appears to be lack of studies examining the multidimensional structure characterising CPS competence and the external aspect of validity concerning measures derived from computer-based assessments of CPS competence. This paper aims to contribute to the validation of the PISA 2015 CPS competence measure using a unified validity definition (Messick, 1995) and applying the Rasch measurement framework (Rasch, 1960). The following research question and sub-questions guided this paper:

> RQ3. What are the strengths and limitations of measurement validity of the CPS competence measure for England based on PISA 2015 data?
>
>> RQ3.a: To what extent is a hypothetical three-dimensional structure of 'establishing and maintaining shared understanding', 'taking appropriate action to solve the problem', and 'establishing and maintaining team organisation' measures supported empirically by the PISA 2015 data for CPS assessment in England?
>>
>> RQ3.b: To what extent are the constructed measures of CPS competence invariant across gender?
>>
>> RQ3.c: How are the constructed measures of CPS competence related to other relevant (collaboration and performance) constructs?

6.4 Methods

6.4.1 Data

Publicly available student-level data from the PISA 2015 study (available at: https://www.oecd.org/pisa/data/) are used. The PISA study is conducted every 3 years, starting from 2000, and assesses students’ proficiency in science, mathematics and reading. The target population is 15-year-olds in schools. The PISA study employs a two-stage sample design, with schools selected first with probability proportional-to-size, and students randomly selected from each school. In 2015 cycle, science was the major domain of assessment and CPS was added for the first time as an innovative domain (OECD, 2017a). Due to time constraints[21], students were randomly assigned a subsample of test questions (clusters of items) to complete. Consequently, although all students answered questions in science (i.e., major domain), only around 30% of students answered any questions in CPS (OECD, 2017c, p. 40). To address the research questions, PISA 2015 data for England[22] are used, with a total of 1,584 students being administered at least one cluster of CPS items (52.34% male and 47.66% female). For more details regarding the PISA 2015 test design and sampling approach see Chapter 3.

6.4.2 PISA 2015 collaborative problem-solving competence assessment design

The PISA 2015 CPS assessment comprises a total of 117 test questions (items) distributed across six computer-simulated, scenario-based tasks, also called assessment units. Each task involved a scenario with multiple individual items (12 to 28 items) and required between 5 and 20 minutes to complete. The CPS tasks were designed to capture the competence of individuals to work in collaborative settings and this was achieved by having students interact with pre-programmed computer-simulated partners instead of other humans. To complete the task, students had to make a multiple-choice selection of ways to respond through predefined messages presented to them in a chat area, or to perform actions (e.g., dragging and dropping) in the visual display area (OECD, 2017b). No free-response items were

[21] PISA is a 2-hour test testing multiple different subjects.

[22] In England, a total of 5,194 students from 206 schools participated in PISA 2015.

available to students. An example of an item used in PISA 2015 CPS assessment is presented in Chapter 3.

Each item was designed to measure one CPS skill and was coded in two (dichotomous: 0/1) or more (polytomous: 0, 1, . . . 4) categories. The extracted data consists of 97 dichotomous items and 20 polytomous items. Depending on students' responses, they received full credit, partial credit, or no credit, which refer to correct, partially correct, and incorrect response categories respectively. More precise details on the scoring of each item can be found in Appendix 4.

6.4.3 Analytical approach

Analysis involves a sequential two-step procedure: the construct validation, and the modelling with constructed measures.

In **step one**, students' responses to the PISA 2015 CPS assessment are used and a series of Rasch measurement models are employed. First, the Partial Credit model is used to analyse all CPS items, assuming they measure a single construct (i.e., CPS competence), ignoring any distinctions between possible underlying dimensions. Second, the Partial Credit model is used to analyse separately the three hypothesised dimensions that are based on the different CPS competencies that the items target (see also Table 6.1). The following nomenclature is used to describe the three dimensions throughout the remaining sections: "Shared understanding" denoted the dimension "Establishing and maintaining shared understanding"; "Taking actions" denoted the dimension "Taking appropriate action to solve the problem"; and "Team organisation" denoted the dimension "Establishing and maintaining team organisation". Each hypothesised dimension is modelled as a unidimensional construct producing three separate student estimates.

In **step two**, the constructed CPS competence measures are used as variables in further statistical analyses, including regression modelling, to evaluate external and consequential aspects of validity. Specifically, correlations between sets of measures are used to report similarities between related constructs (i.e., attitudes towards collaboration and CPS competence measures). Regression modelling is also utilised to determine whether (theory-based) predictions about changes within individuals are realised in the measures of CPS competence. The validation process is conducted within the Rasch measurement framework and follows widely accepted Rasch guidelines (Wolfe & Smith, 2007a, 2007b), which are in turn based on Messick's (1989) definition of validity. In the following sections, the analysis conducted in these two steps is presented briefly, for a more detailed description see Chapter 3.

6.4.4 Measurement approach to validation (Step 1)

6.4.4.1 Partial credit model

The Partial Credit model (Masters, 1982; Wright & Masters, 1982), which is an extension of the dichotomous Rasch model (Rasch, 1960), is used to construct and validate the measures of CPS competence. The Partial Credit model is appropriate for items scored either dichotomously or polytomously by awarding partial credit for responses that are neither correct nor totally incorrect. For the Partial Credit model, additional parameters are added to the Rasch model: the difficulty of achieving each of the (ordered) score categories.

As in the Rasch model, greater 'ability' corresponds to a higher probability of achieving a larger score on an item. This process of internal validation with the Partial Credit model, follows guidelines from Wolfe and Smith (2007a, 2007b) but also Bond and Fox (2007), and Messick's (1995) validity framework concerning content, structural, generalisability, external, and consequential validity. Decisions about validity are informed by different statistical indices, such as item fit statistics, dimensionality diagnostics, reliability and separation statistics, person-item maps, and differential item functioning (Bond & Fox, 2007). The software package WINSTEPS (Linacre, 2006c) is used for item calibration.

6.4.4.2 Item fit statistics

Item fit statistics indicate how accurately the data fit the model providing evidence in support (or not) of the unidimensionality assumption. The ideal expected value is close to 1.0. Considering existing guidelines (Adams & Khoo, 1995; Bond & Fox, 2007) and previous research (Pampaka, 2021; Pampaka et al., 2013), I consider values above 1.3 to suggest causes for concern. Standardised fit statistics (Zstd) are t-tests of the hypothesis that data fit the model perfectly. A Zstd value should be flagged as significant if the absolute value is larger than 1.96. Residual correlations for items are also examined to check for items that may be locally dependent. High positive residual correlations (around 0.70) may indicate local item dependency between pairs of items.

6.4.4.3 Principal components analysis of residuals

Evidence relating to the structural aspect of validity is explored through the results of a principal component analysis of the residuals, which can aid in determining whether the measure under investigation approximates a unidimensional measure (Linacre, 1998; Smith, 2002; Wolfe & Smith, 2007b). If the eigenvalue of the first contrast (i.e., the component that explains the largest possible amount of variance in the residuals) is small, usually less than two units, then the first contrast is at the noise level (Linacre, 2006a).

6.4.4.4 Reliability and separation

Person separation indicates how efficiently a set of items can separate those persons measured. Item separation indicates how well a sample of people can separate those items used in the test (Wright & Stone, 1999). Where these statistics are expressed as reliabilities, they range from 0.0 to 1.0, the higher the value the better the separation that exists and the more precise the measurement (Wright & Stone, 1999).

6.4.4.5 Person-item map and item characteristic curve plots

A person-item map (also called Wright map) visualises the location of the item difficulties and the distribution of student 'abilities'. If the distribution of student 'ability' is skewed as compared to the item difficulty distribution, then more items might be needed to capture appropriately the construct. Additionally, model fit can be visualised through item characteristic curve plots, which show the relationship between student latent 'ability' and the probability of a response to a given category within an item.

6.4.4.6 Establishing measurement invariance – differential item functioning analysis

Differential item functioning (DIF) analysis seeks to determine whether two groups have different probabilities of providing a particular response to individual items, when matched on measures of the construct (Wolfe & Smith, 2007b). This paper reports on DIF by gender, following relevant guidelines (Zwick, 2012) for DIF size and its statistical significance.

6.4.5 Statistical modelling with constructed measures (Step 2)

Following construct validation, the derived scores (i.e., students' measures) are added to the original dataset along with the other student variables for further modelling.

6.4.5.1 Variables

CPS competence measures: overall and sub-scales. The resulting person (student) scores of the four constructed CPS competence measures (overall score[23] and three sub-scale scores), are the main outcome variables used throughout the analyses.

Attitudes towards collaboration. As part of the PISA 2015 background questionnaire, students were asked eight questions about their attitudes towards

[23] The correlation between the overall CPS competence scale (Rasch student scores) and the PISA's student CPS performance (10 plausible values) was 0.94, p<0.001.

collaboration and two derived variables are provided in the dataset. The first variable, valuing relationships, is related to altruistic interactions, when the student engages in collaborative activities. The second, valuing teamwork, is related to what teamwork can produce, as opposed to working alone. Items were scored so as higher values correspond to more positive attitudes towards collaboration. The two scales' reliabilities (Cronbach's Alpha) were 0.723 and 0.821[24], respectively (OECD, 2017c, p. 307).

Science, mathematics, and reading performance. A set of ten plausible values were drawn for each pupil in each subject area tested in PISA 2015. These have a mean of around 500 points and a standard deviation of around 100 points. I used the ten plausible values included in the PISA 2015 dataset for each subject (i.e., science, mathematics, reading). Following recommended practice (OECD, 2009), each model was estimated ten times (once for each plausible value) with the parameter estimates and standard errors then pooled according to 'Rubin's rules' (Rubin, 1987). The Stata 'REPEST' package, developed by members of the OECD (Avvisati & Keslair, 2014), was used for this analysis.

Economic, social, and cultural status (ESCS) index. The PISA ESCS index is a composite score based on three other indices reflecting parental education, highest parental occupation, and home possessions built via principal component analysis. The reliability (Cronbach's Alpha) of the scale was 0.63[25] (OECD, 2017c, p. 340).

Gender. The dummy variable 'female' with assigned values: 0 = male and 1 = female.

[24] Scale reliabilities correspond to sample from the United Kingdom as reported in the OECD technical report.

[25] Scale reliability corresponds to sample from the United Kingdom as reported in the OECD technical report (OECD, 2017c).

Geographic region. This categorical variable describes the location of participant's school, with assigned values: 1 = Greater London, 2 = South, 3 = Midlands, and 4 = North.

School type. This categorical variable describes the type of participant's school, with assigned values: 1 = academy, 2 = maintained selective, 3 = maintained non-selective, and 4 = independent.

Descriptive statistics of all the above variables are presented in Appendix 5.

6.4.5.2 Missing values

To keep the sample size consistent across the models in further statistical modelling, only cases with complete information were used. The small number of students who had missing values in the variables of interest (around 6%) were excluded, resulting in an analytical sample size of 1,485 cases. More detail on missing values by variables can be found in Appendix 6.

6.4.5.3 Weights

To account for the PISA study complex survey design (stratified and clustered sample design), final student weights and balanced-repeated-replication weights are applied throughout the analysis. The balanced-repeated-replication weights are based upon a resampling method and allow the impact of both the stratification and clustering to be incorporated into the estimated standard error (Jerrim et al., 2019). These weights are provided with the data and are recommended by the survey organisers (OECD, 2009). The Stata version 16 (StataCorp, 2019) and the Stata 'REPEST' package (Avvisati & Keslair, 2014) are used to apply the above weights and conduct data analysis.

6.4.5.4 Analyses

To explore the associations of personal features and the CPS competence measures (overall and sub-scales), the resulting scores of these measures were initially compared based on gender. Secondly, correlation analysis was conducted between CPS competence scores and student attitudes towards collaboration as well as subject performance. Following that, regression analyses were conducted to model the relationship between personal, contextual, attitudinal, and performance features on students' CPS competence measures (overall and sub-scales). For regression analyses, all continuous (dependent and independent) variables were standardised to have mean zero and a standard deviation of one across the population of interest.

Regression models were run in the following order to illustrate how parameter estimates changed with the addition of extra variables. Each model was run four times using each CPS competence measure (overall and sub-scales) as outcome variable:

- Model 1a-1d: only basic demographic characteristics.
- Model 2a-2d: Model 1 plus students' attitudes towards collaboration. A set of dummy variables referring to quartiles of the attitudes towards collaboration scales were entered into the models. The bottom quartile (negative attitudes) was set as the reference group in both scales.
- Model 3a-3d: Model 2 plus students' science performance[26] (10 plausible values).

[26] PISA science score was added as independent variable in regression analyses since it was the major subject of assessment in the PISA 2015 cycle and to avoid multicollinearity issues.

6.5 Results

6.5.1 Validation results (Step 1)

6.5.1.1 Overall collaborative problem-solving competence measure

Item calibration combined all items (n = 117) assuming they measure a single construct. All weighted fit values are within the adopted acceptable range of 0.70-1.30, apart from one item with INFIT value 1.33 (item 48). For almost half of the items, the standardised values are bigger than |2|, however p-values are usually high with big sample sizes. Most of the un-weighted fit values are within the acceptable range, apart from a small number of items (n = 11). All item fit statistics for the constructed measure are provided in Appendix 17. Results concerning separation and reliability are as follows:

- Item separation = 12.26, reliability = 0.99
- Person separation = 2.70, reliability = 0.88

This shows good person separation suggesting that the test discriminates the sample into two or three levels of 'ability' and good item separation, which implies that the person sample is large enough to precisely locate the items on the latent variable (Linacre, 2006b). The residual correlations for items were found to be low (<0.70), suggesting no issues about dependency.

Figure 6.1 illustrates the item characteristic curve plots of two example items: a good fitting dichotomous item (item 80) and a poorer fitting partial credit item (item 48) with four possible response categories. For item 48, Figure 6.1 shows that misfit is present at the low end of the 'ability' continuum (under-discrimination, fit statistic 1.33) and that that the two (middle) scores are not most likely to be used at any point of measure distribution, which explains the misfit there. If the item content was released, this item could be further reviewed to understand more about the measure and the impact of test design on the item quality.

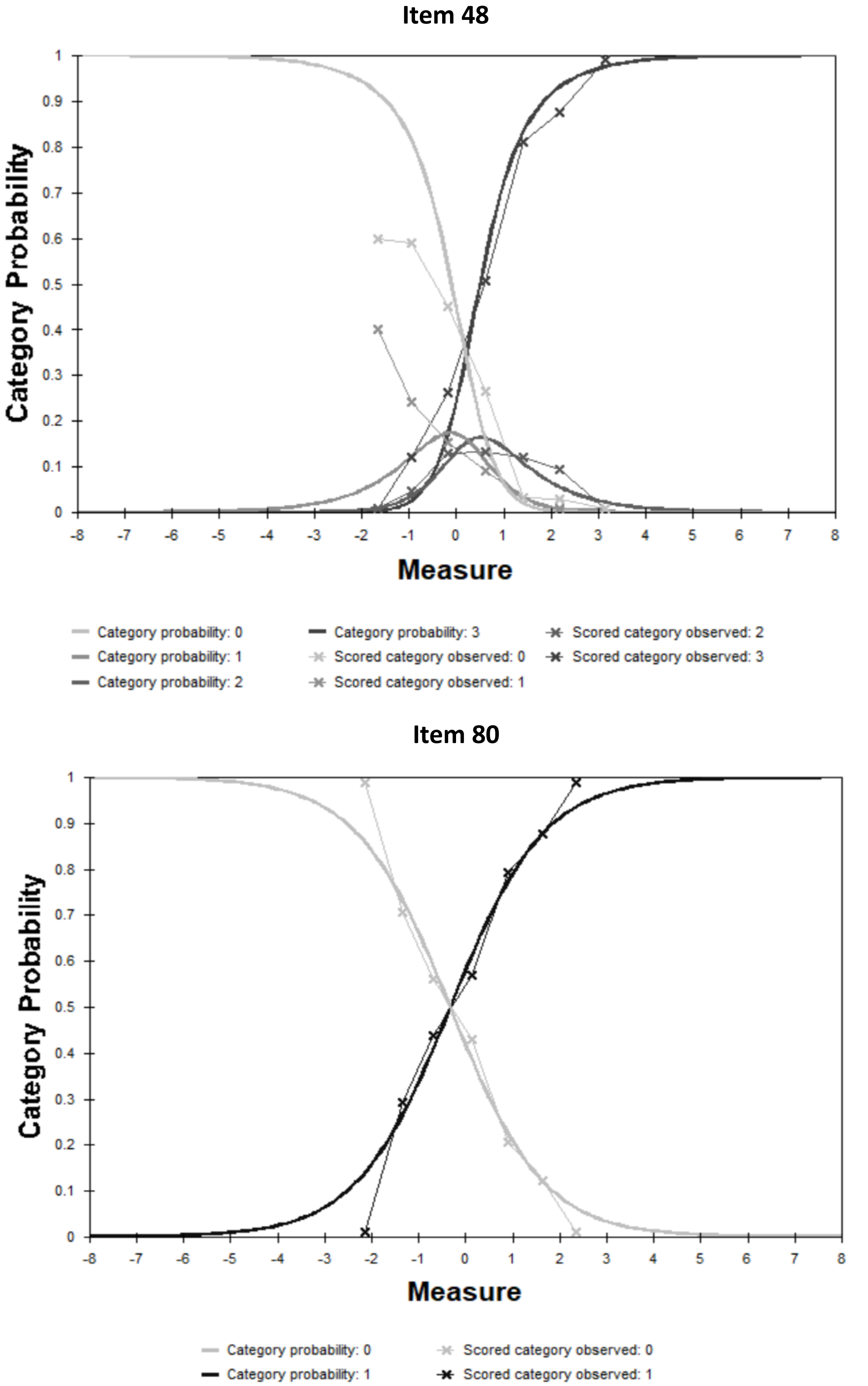
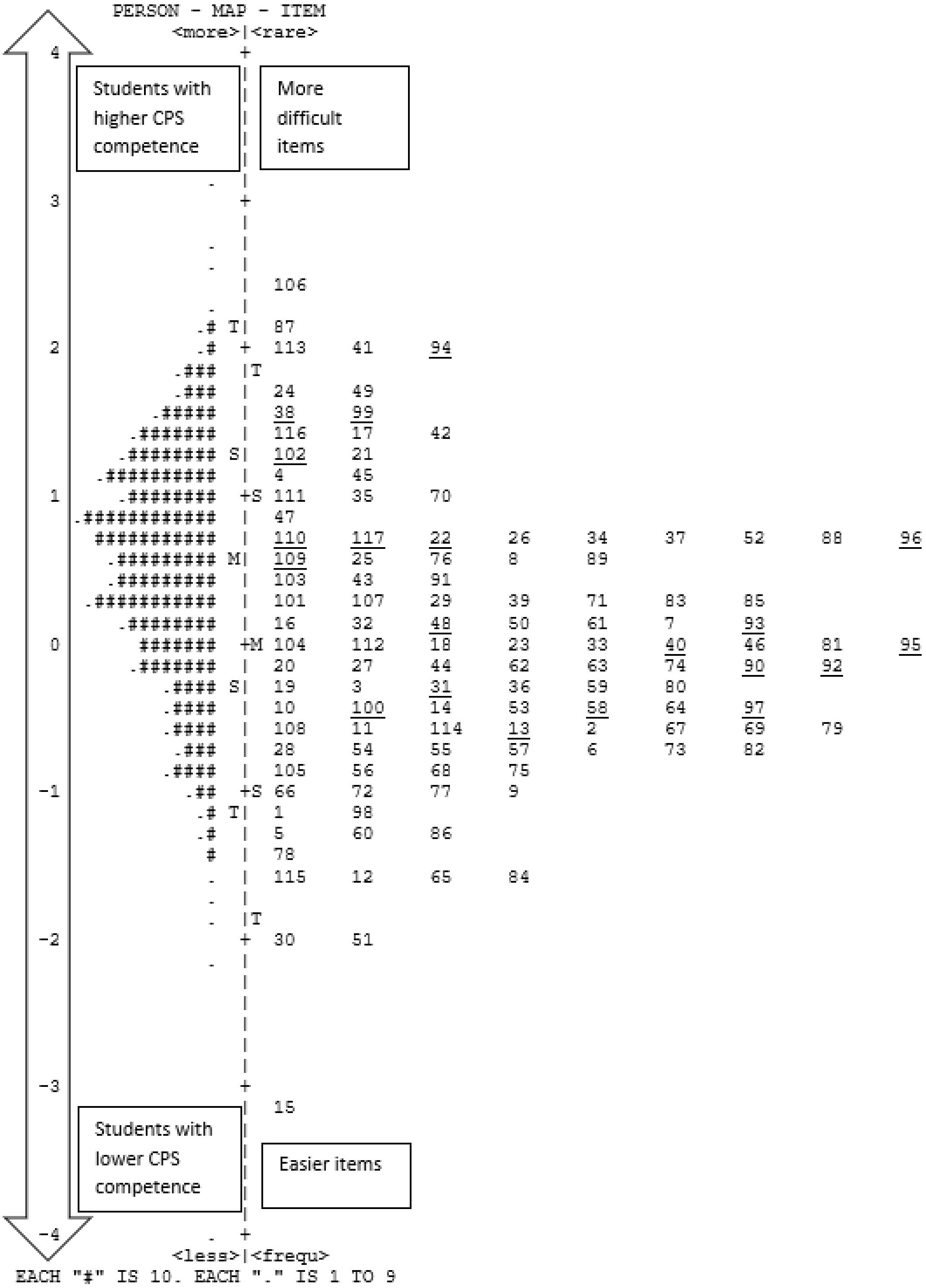


Figure 6.1. Item characteristic curve plots showing a poorer fitting item (top) and a better fitting item (bottom) from the overall CPS competence scale.

The person-item map, presented in Figure 6.2, shows the distribution of items based on their difficulty to be endorsed by students. The logit scale (denoted with the numbers ranging from -4 to +4) is the common measurement scale for both items and persons (i.e., students). Students at the bottom of the scale score low, while students on the top end of the scale score high on the CPS competence scale. The plot shows that, overall, there is a good coverage of items along the scale, with most items located around the middle of the scale. Most students found most of the items easy (to get right), as they (the students) are mainly located at the top half of the scale. It should be noted that for some items (n = 20) the location is the average difficulty, since they are polytomous (see Figure 6.2, items indicated with underline).

The four most difficult items to get right (items 41, 87, 106 and 113), located at the top of the scale, target “Establishing and maintaining shared understanding”. Content of the released item 87 needs to be reviewed to better understand the high item difficulty. The difficulty of the item could be attributed to the fact that students needed to track not only the chat space, but also notice a change in the problem space, which violated the previously agreed rules of engagement (OECD, n.d.). In addition, it could be argued that the credited response did not sound overtly collaborative, and therefore, it could be speculated that students avoided it due to its tone. However, there is no other information available to help interpretation of that result. Specifically, descriptions of students’ possible actions in the task were not available in the data set.



Students with higher CPS competence

More difficult items

Students with lower CPS competence

Easier items

```
             PERSON - MAP - ITEM
                  <more>|<rare>
  4                     +
                        |
                        |
                        |
                        |
                        |
                      . |
  3                     +
                        |
                      . |
                      . |
                        |  106
                      . |
                   .# T|  87
  2                .#   +  113    41     94
                .###    |T
                .###    |  24     49
              .#####    |  38     99
            .#######    |  116    17     42
           .######## S|  102    21
         .##########    |  4      45
  1        .########   +S  111    35     70
      .############    |  47
       ###########     |  110    117    22     26     34     37     52     88     96
        .######### M|  109    25     76     8      89
        .#########     |  103    43     91
      .###########     |  101    107    29     39     71     83     85
         .########     |  16     32     48     50     61     7      93
  0        #######    +M  104    112    18     23     33     40     46     81     95
          .#######    |  20     27     44     62     63     74     90     92
             .#### S|  19     3      31     36     59     80
             .####    |  10     100    14     53     58     64     97
             .####    |  108    11     114    13     2      67     69     79
              .###    |  28     54     55     57     6      73     82
             .####    |  105    56     68     75
 -1            .##   +S  66     72     77     9
                .# T|  1      98
                .#    |  5      60     86
                 #    |  78
                 .    |  115    12     65     84
                 .    |
                 .    |T
 -2                   +  30     51
                 .    |
                      |
                      |
                      |
                      |
                      |
 -3                   +
                      |  15
                      |
                      |
                      |
                      |
                      |
 -4              .    +
                  <less>|<frequ>
 EACH "#" IS 10. EACH "." IS 1 TO 9
```

Figure 6.2. Person-item map for the overall CPS competence scale including 117 items

Notes: Items with underline have three or more response categories.

The results of the principal components analysis of residuals are shown in Table 6.2. As can be seen from the table, the Rasch dimension explains 32.7% of the variance in the data, much bigger than the variance explained by the first contrast in the residuals. However, the eigenvalue for unexplained variance in the first contrast has a size of 2.8, that is the strength of about three items (i.e., the smallest amount that could be considered a "dimension"). It is therefore suggested that, although its strength is quite small, there may be issues with constructs dimensionality which might suggest the existence of up to three useful sub-scales.

Table 6.2. Standardised residual variance (in Eigenvalue units)

	Empirical			Modeled
Total raw variance in observations	173.9	100.0%		100.0%
Raw variance explained by measures	56.9	32.7%		34.4%
Raw variance explained by persons	28.5	16.4%		17.2%
Raw Variance explained by items	28.4	16.3%		17.2%
Raw unexplained variance (total)	117.0	67.3%	100.0%	65.6%
Unexplained variance in 1st contrast	2.8	1.6%	2.4%	
Unexplained variance in 2nd contrast	2.0	1.2%	1.7%	

Figure 6.3 illustrates results from DIF analysis conducted to account for possible instrument bias. Item 12 appears to have a big value of DIF size (bigger than 0.64 logit difference), while a few more items, such as 54 and 70, show moderate DIF size (from 0.43 to 0.64 logit difference). Item content is not available for these items, and therefore it is not possible to explain whether some features of these items are responsible for the DIF results. Overall, the measure appears to be invariant by gender, with a few exceptions, i.e., items showing a DIF size bigger than 0.43 (Zwick, 2012).

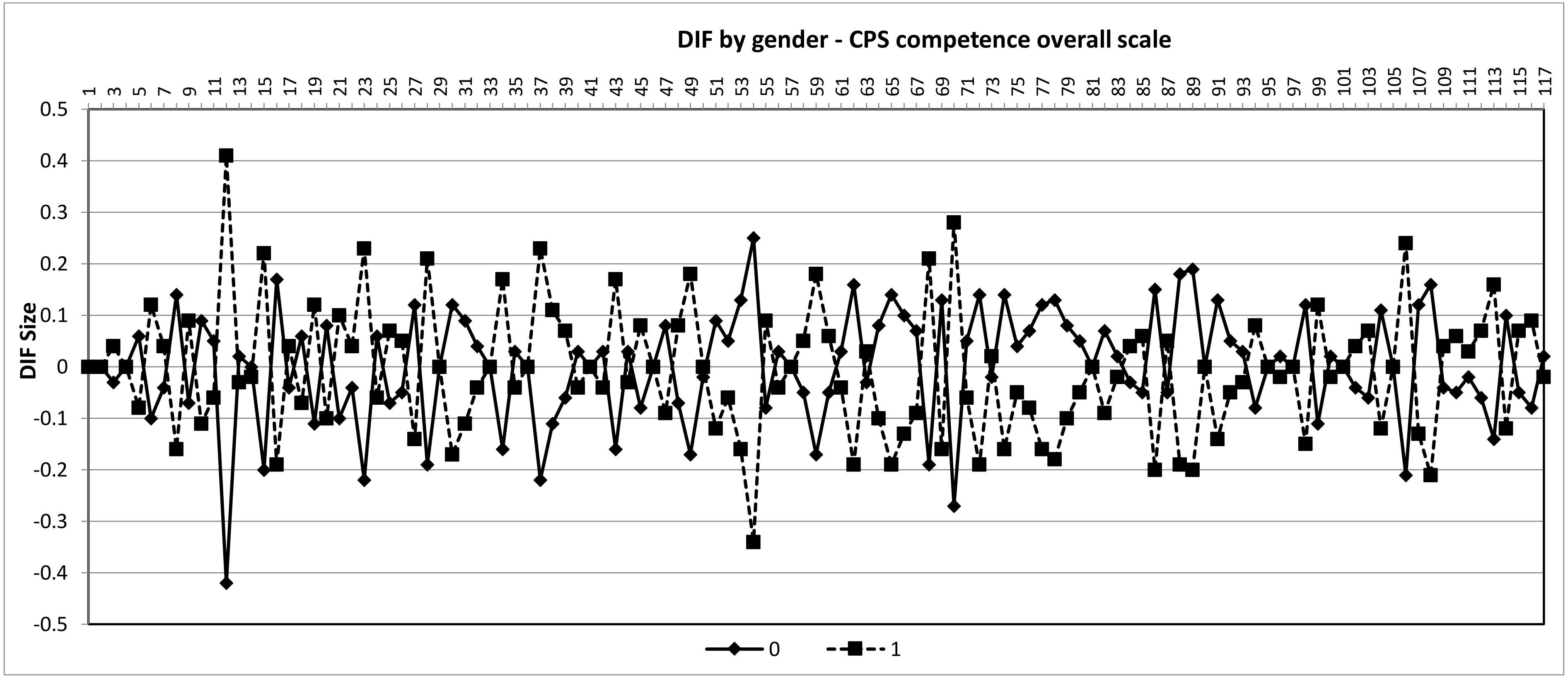


Figure 6.3. Person fit plot by gender (0 = male, 1 = female) – Scale “CPS competence overall” with 117 items.

Notes: Analytical sample: n = 1,584 students (52.34% male and 47.66% female).

6.5.1.2 Three sub-scales

Guided by the multi-dimensional structure of the PISA 2015 CPS theoretical framework, three models were further examined, each representing one of the three hypothesised sub-scales: “Shared understanding”, “Taking actions”, and “Team organisation”. The same procedure is followed for these additional constructed measures. The three possible sub-scales show a good overall functionality, but some weaknesses compared to the overall CPS competence scale. Specifically, all INFIT values were close to the ideal value of 1.0, thus confirming a good fit of the data to the three models. Item fit statistics for the three constructed measures are provided in Appendix 18.

Results concerning separation and reliability for the three sub-scale measures are as follows:

- “Shared understanding”: Item separation = 11.63, reliability = 0.99;
  Person separation = 1.98, reliability = 0.80
- “Taking actions”: Item separation = 13.94, reliability = 0.99;
  Person separation = 1.27, reliability = 0.62
- “Team organisation”: Item separation = 12.76, reliability = 0.99;
  Person separation = 1.19, reliability = 0.58

The three sub-scales showed high item separation and item reliability. Nonetheless, “Taking actions” and “Team organisation”, were found to have lower person separation and person reliabilities than the ideal values. This is not unexpected though, since there was a smaller set of items in those two sub-scales, compared to “Shared understanding”, which showed adequate person separation statistics. To strengthen the sensitivity of the two sub-scales, more items targeting “Taking actions” and “Team organisation” could be added in the test. The residual correlations for items were found to be low (<0.70), suggesting no issues about dependency.

The three measures appear to be invariant by gender, with very few exceptions, i.e., items showing DIF size bigger than 0.43 (results of DIF analysis presented in Appendix 19). None of the items showing moderate DIF size are included in the visible CPS task Xandar, and therefore, further interpretation of DIF analysis results is difficult to be pursued.

Figure 6.4 shows the person-item maps for the three sub-scales. As explained earlier, each hypothesised sub-scale was measured by a different set of items. Furthermore, the results of the three corresponding principal components analysis of residuals were examined (see details in Appendix 20). For "Taking actions" and "Team organisation", all the eigenvalues for unexplained variance in additional contrasts are smaller than 2 (suggesting that there are no serious issues with the constructs' unidimensionality). For "Shared understanding", the eigenvalue for unexplained variance in the first contrast has a size of 2.1, which is still quite small (this suggests that the first contrast is at the noise level). The construct validation step is concluded with the validation of the CPS competence measures (one overall and three sub-scales). In the following section, results from the statistical analysis, conducted using the derived scores (i.e., Rasch person scores), are presented.

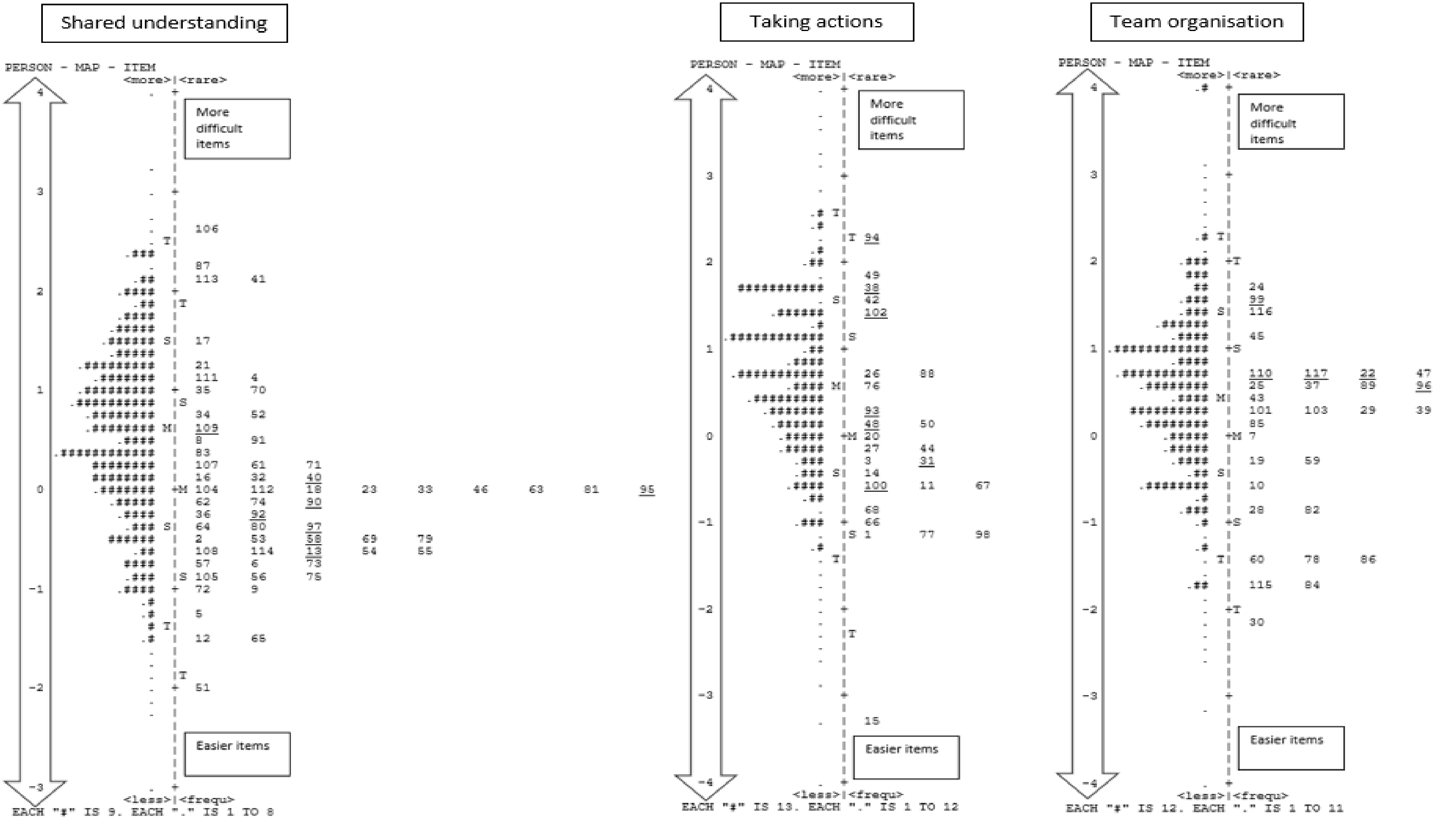


Figure 6.4. Person-item maps for the three sub-scales: “Shared understanding”, “Taking actions”, and “Team organisation”

6.5.2 Statistical modelling with constructed measures (Step 2)

6.5.2.1 Descriptive statistics

Table 6.3 shows the mean comparisons of CPS competence Rasch test scores (overall and sub-scales) by gender. Considering that higher Rasch scores indicate higher CPS competence, significant gender differences in favour of girls were observed in CPS competence scores (overall and sub-scales). The biggest difference between girls and boys (0.30 points) appeared in “Taking actions”.

Table 6.3. Mean comparisons of CPS competence scores by gender

	Male (n = 771)	Female (n = 714)		
	Mean (SD)	Mean (SD)	t-test	p
CPS competence (Overall)	0.36 (0.82)	0.59 (0.79)	-4.85	0.00
Sub-scale 1 (Shared understanding)	0.40 (0.95)	0.67 (0.94)	-5.06	0.00
Sub-scale 2 (Taking actions)	0.36 (1.05)	0.66 (0.99)	-4.20	0.00
Sub-scale 3 (Team organisation)	0.34 (0.98)	0.50 (0.96)	-2.86	0.01

Notes: Analytical sample (N=1,485)

6.5.2.2 Correlation analysis

Table 6.4 provides results from correlation analysis. Panel (a) presents the relationship between the CPS competence measures (overall and sub-scales) with students’ attitudes towards collaboration and subject performance, and Panel (b) provides the correlation between the three CPS competence sub-scales. Interestingly, the correlations between CPS competence measures and students’ attitudes towards collaboration are quite low ($r < |0.18|$). The direction of the relationship is opposite for the two attitudinal scales; valuing teamwork is negatively associated, while valuing relationships is positively associated with CPS competence measures.

As shown in Table 6.4, the strongest relationship can be observed between CPS competence overall measure and science performance with r = 0.74, while the correlation coefficients of CPS competence sub-scales and science performance ranged from 0.61 to 0.70. The correlations between CPS competence overall measure and the remaining two subjects were also high (r = 0.70 for reading and r = 0.65 for mathematics), while the correlation coefficients between CPS competence sub-scales and the two subjects ranged from 0.52 to 0.66. Finally, the correlations between CPS competence sub-scales were high, with the coefficients ranging from 0.61 to 0.70.

Table 6.4. Pearson correlation between students' CPS competence, attitudes, and subject performance

(a) Correlation between the CPS competence (overall and sub-scales) and other constructs

	Valuing teamwork	Valuing relationships	Maths	Science	Reading
CPS competence *(Overall)*	-0.08**	0.18***	0.65***	0.74***	0.70***
Sub-scale 1 *(Shared understanding)*	-0.08**	0.17***	0.62***	0.70***	0.66***
Sub-scale 2 *(Taking actions)*	-0.07*	0.13***	0.55***	0.63***	0.60***
Sub-scale 3 *(Team organisation)*	-0.06*	0.16***	0.52***	0.61***	0.57***

(b) Correlation between the CPS competence sub-scales

	Sub-scale 1 *(Shared understanding)*	Sub-scale 2 *(Taking actions)*	Sub-scale 3 *(Team organisation)*
Sub-scale 1 *(Shared understanding)*	1		
Sub-scale 2 *(Taking actions)*	0.70***	1	
Sub-scale 3 *(Team organisation)*	0.70***	0.61***	1

Notes: * p<0.05, ** p<0.01, *** p<0.001, Correlations above 0.5 shaded in grey.

6.5.2.3 Regression modelling

Table 6.5 presents estimates from the first regression model, which includes basic demographic characteristics (Model 1a – 1d). These results suggest that being a girl is associated with a 0.29-point increase in overall CPS competence, “Shared understanding”, and “Taking actions” scores, as compared to being a boy. For “Team organisation” scores, girls still perform better than boys, but the increase in their scores is somewhat smaller (0.17). Economic, social, and cultural status is positively associated with student scores in all CPS competence scales, while geographic region of students’ school is not significantly associated with their CPS competence. For all outcome measures, being in a maintained selective school is positively associated with students’ scores, as compared to being in an academy. On the contrary, being in a maintained non-selective school is negatively associated with students’ CPS competence scores (overall, shared understand, and team organisation), as compared to being in an academy.

Table 6.5. Regression model estimates: Model 1a-1d

	Model 1a (Overall CPS competence)			**Model 1b** (Shared understanding)			**Model 1c** (Taking actions)			**Model 1d** (Team organisation)		
Variables	b	S.E.	p	b	S.E.	p	b	S.E.	p	b	S.E.	p
Gender (ref. Male)												
Female	0.29	0.06	0.00	0.29	0.05	0.00	0.29	0.07	0.00	0.17	0.06	0.00
ESCS	0.24	0.03	0.00	0.24	0.03	0.00	0.20	0.03	0.00	0.20	0.03	0.00
Region (ref. Greater London)												
South	0.05	0.11	0.61	-0.02	0.11	0.85	0.15	0.10	0.15	0.04	0.09	0.66
Midlands	-0.04	0.10	0.69	-0.08	0.11	0.48	0.01	0.09	0.92	0.04	0.09	0.69
North	-0.01	0.11	0.94	-0.03	0.11	0.81	0.02	0.09	0.86	0.03	0.09	0.71
School type (ref. academy)												
Maintained selective	0.53	0.07	0.00	0.49	0.07	0.00	0.40	0.10	0.00	0.48	0.05	0.00
Maintained non-selective	-0.15	0.06	0.02	-0.14	0.07	0.03	-0.12	0.06	0.05	-0.15	0.06	0.01
Independent	0.03	0.11	0.80	0.02	0.11	0.89	0.08	0.09	0.38	-0.01	0.10	0.92
Constant	-0.09	0.10	0.36	-0.06	0.11	0.58	-0.16	0.09	0.08	0.04	0.08	0.64
R squared	0.09	0.02	0.00	0.09	0.02	0.00	0.08	0.02	0.00	0.06	0.01	0.00
Observations	1,485			1,485			1,485			1,485		

Notes: Estimates refer to the increase in CPS competence scores (overall, shared understanding, taking actions, team organisation). Grey shading indicates statistically significant at the 0.05 level and SE = standard errors. The number of students in each quartile of the valuing teamwork scale was 381 (bottom quartile), 627 (second quartile), 129 (third quartile), and 348 (top quartile). The number of students in each quartile of the valuing relationships scale was 394 (bottom quartile), 544 (second quartile), 219 (third quartile), and 328 (top quartile).

Students’ attitudes towards collaboration were added in Models 2a-2d. As shown in Table 6.6, the addition of the two attitudinal measures as dummy variables does not significantly change background variables’ associations with CPS competence measures. Interestingly, valuing relationships and valuing teamwork have opposite association with the CPS competence measures. For valuing teamwork, students with very positive attitudes towards teamwork (top quartile) have a 0.49-point decrease in their overall CPS competence scores compared to those with very negative attitudes towards teamwork (bottom quartile). There is also a negative association when comparing the second quartile versus bottom quartile groups. Similar results can be observed for the association of valuing teamwork quartiles and the three CPS competence sub-scales.

For valuing relationships, students with very positive attitudes (top quartile) have a 0.58-point increase in their overall CPS competence scores, compared to students who have very negative attitudes towards relationships (bottom quartile). A positive association, although somewhat smaller, appears when comparing third and second quartile versus bottom quartile groups. For “Shared understanding”, the association between the valuing relationships’ quartiles and students’ scores is similar to the overall CPS competence measure. For “Taking actions”, students with positive and very positive attitudes in valuing relationships have higher scores compared to students with very negative attitudes. Finally, for “Team organisation”, students with negative and very positive attitudes in valuing relationships have higher scores compared to students with very negative attitudes.

Table 6.6. Regression model estimates: Model 2a-2d

	Model 2a (Overall CPS competence)			**Model 2b** (Shared understanding)			**Model 2c** (Taking actions)			**Model 2d** (Team organisation)		
Variables	b	S.E.	p	b	S.E.	p	b	S.E.	p	b	S.E.	p
Gender (ref. Male)												
Female	0.25	0.06	0.00	0.25	0.05	0.00	0.26	0.07	0.00	0.13	0.06	0.02
ESCS	0.22	0.03	0.00	0.21	0.03	0.00	0.18	0.03	0.00	0.18	0.03	0.00
Region (ref. Greater London)												
South	0.10	0.10	0.35	0.02	0.11	0.89	0.19	0.10	0.06	0.08	0.09	0.36
Midlands	0.01	0.10	0.94	-0.03	0.11	0.76	0.05	0.09	0.57	0.01	0.09	0.91
North	0.00	0.10	0.99	-0.02	0.11	0.83	0.02	0.09	0.79	0.04	0.09	0.64
School type (ref. academy)												
Maintained selective	0.52	0.07	0.00	0.48	0.07	0.00	0.39	0.11	0.00	0.48	0.06	0.00
Maintained non-selective	-0.15	0.06	0.01	-0.14	0.07	0.03	-0.12	0.06	0.04	-0.15	0.06	0.01
Independent	0.04	0.10	0.67	0.02	0.10	0.83	0.08	0.08	0.32	0.02	0.09	0.82
Valuing teamwork (ref. bottom quartile)												
Second quartile	-0.29	0.07	0.00	-0.29	0.07	0.00	-0.22	0.06	0.00	-0.24	0.07	0.00
Third quartile	-0.15	0.12	0.23	-0.16	0.12	0.20	-0.17	0.13	0.19	-0.10	0.12	0.42
Top quartile	-0.49	0.09	0.00	-0.46	0.09	0.00	-0.39	0.08	0.00	-0.44	0.10	0.00
Valuing relationships (ref. bottom quartile)												
Second quartile	0.21	0.07	0.00	0.23	0.07	0.00	0.12	0.07	0.09	0.17	0.07	0.01

Third quartile	0.29	0.11	0.01	0.34	0.12	0.01	0.21	0.10	0.03	0.15	0.10	0.13
Top quartile	0.58	0.07	0.00	0.55	0.08	0.00	0.44	0.08	0.00	0.53	0.08	0.00
Constant	-0.10	0.12	0.43	-0.07	0.13	0.58	-0.13	0.11	0.22	-0.04	0.11	0.73
R squared	0.14	0.02	0.00	0.13	0.02	0.00	0.10	0.02	0.00	0.10	0.02	0.00
Observations	1,485			1,485			1,485			1,485		

Notes: Estimates refer to the increase in CPS competence scores (overall, shared understanding, taking actions, team organisation). Grey shading indicates statistically significant at the 0.05 level and SE = standard errors. The number of students in each quartile of the valuing teamwork scale was 381 (bottom quartile), 627 (second quartile), 129 (third quartile), and 348 (top quartile). The number of students in each quartile of the valuing relationships scale was 394 (bottom quartile), 544 (second quartile), 219 (third quartile), and 328 (top quartile).

When adding science performance score in Model 3a-3d (Table 6.7), girls still have a 0.30-point increase in their CPS competence overall, “Shared understanding”, and “Taking actions” scores, and a 0.17-point increase in their “Team organisation” scores, as compared to boys. Economic, social, and cultural status does not reach statistical significance. The magnitude of the association between attitudes towards collaboration and CPS competence measures drops by more than 50% and does not reach statistical significance. The only exception is observed in the “Team organisation” measure; students with very positive attitudes in valuing relationships have higher scores as compared to those with very negative attitudes. Finally, results from Table 6.7 suggest that there is a strong positive association between CPS competence scores (overall and sub-scales) and students’ PISA scores in science, e.g., a unit of increase in science scores is associated with a 0.74-point increase in overall CPS competence scores.

Table 6.7. Regression model estimates: Model 3a-3d

	Model 3a (Overall CPS competence)			**Model 3b** (Shared understanding)			**Model 3c** (Taking actions)			**Model 3d** (Team organisation)		
Variables	b	S.E.	p	b	S.E.	p	b	S.E.	p	b	S.E.	p
Gender (ref. Male)												
Female	0.30	0.04	0.00	0.30	0.04	0.00	0.30	0.05	0.00	0.17	0.05	0.00
ESCS	0.02	0.02	0.38	0.02	0.02	0.20	0.01	0.02	0.61	0.01	0.02	0.57
Region (ref. Greater London)												
South	0.01	0.07	0.84	-0.06	0.07	0.40	0.12	0.06	0.07	0.01	0.08	0.86
Midlands	0.05	0.07	0.45	0.01	0.08	0.92	0.09	0.06	0.13	0.05	0.08	0.56
North	0.06	0.07	0.42	0.03	0.08	0.72	0.07	0.06	0.19	0.09	0.08	0.25
School type (ref. academy)												
Maintained selective	0.03	0.06	0.68	0.02	0.05	0.78	-0.03	0.10	0.75	0.07	0.06	0.22
Maintained non-selective	-0.05	0.04	0.24	-0.04	0.05	0.36	-0.03	0.05	0.45	-0.06	0.05	0.19
Independent	-0.16	0.08	0.04	-0.16	0.08	0.04	-0.08	0.07	0.23	-0.14	0.08	0.09
Valuing teamwork (ref. bottom quartile)												
Second quartile	0.02	0.05	0.62	0.00	0.06	0.96	0.05	0.06	0.39	0.02	0.06	0.76
Third quartile	0.09	0.08	0.23	0.07	0.08	0.40	0.04	0.11	0.73	0.10	0.09	0.28
Top quartile	0.03	0.07	0.63	0.03	0.07	0.62	0.05	0.08	0.50	-0.01	0.08	0.90
Valuing relationships (ref. bottom quartile)												
Second quartile	0.02	0.06	0.76	0.05	0.06	0.42	-0.05	0.06	0.39	0.01	0.06	0.83

Third quartile	0.03	0.08	0.69	0.10	0.09	0.27	-0.01	0.07	0.85	-0.06	0.08	0.42
Top quartile	0.14	0.07	0.05	0.14	0.08	0.09	0.07	0.08	0.37	0.17	0.08	0.02
Science performance	0.74	0.02	0.00	0.70	0.02	0.00	0.63	0.03	0.00	0.61	0.03	0.00
Constant	-0.22	0.08	0.01	-0.19	0.09	0.03	-0.24	0.07	0.00	-0.14	0.09	0.13
R squared	0.58	0.02	0.00	0.52	0.02	0.00	0.43	0.02	0.00	0.39	0.02	0.00
Observations	1,485			1,485			1,485			1,485		

Notes: Estimates refer to the increase in CPS competence scores (overall, shared understanding, taking actions, team organisation). Grey shading indicates statistically significant at the 0.05 level and SE = standard errors. The number of students in each quartile of the valuing teamwork scale was 381 (bottom quartile), 627 (second quartile), 129 (third quartile), and 348 (top quartile). The number of students in each quartile of the valuing relationships scale was 394 (bottom quartile), 544 (second quartile), 219 (third quartile), and 328 (top quartile).

6.6 Discussion

In this paper, evidence concerning aspects of validity of CPS competence measures were examined using data from the PISA 2015 study targeting 15-year-olds in England. Results showed that the CPS items collectively measure a unidimensional model, presumed to be of CPS competence. For the sub-scales “Taking actions” and “Team organisation”, more items need to be included in the test to make the measures more accurate. For this set of items/sample, the three sub-scales showed similar results in their associations with other variables analysed, and as compared to the overall CPS competence measure. There are several possible interpretations of a unidimensional measurement model. Unidimensionality could be a result of a common context characterising the assessment of the targeted construct, in this case the computer-simulated, scenario-based environment of the PISA 2015 CPS assessment. Furthermore, it could be that the CPS items assess only a fragment of a more complex CPS competence. Goldstein (2004) has previously argued about the tendency for scales in the PISA study to be constructed as unidimensional, but possibly at the cost of excluding certain kinds of important information that might be discrepant. More specifically, in PISA 2015 CPS assessment, items that exhibited dependencies were combined into polytomous ‘composite items’ to remove local dependencies (OECD, 2017c, pp. 166–170). After the combination into composite items, the number of CPS items available for scaling was reduced.

What is important to consider is the purpose of the assessment and its consequences. For some purposes the unidimensional model is more appropriate, or even demanded, e.g., if there is a need to develop a cut-score below which a person is ‘failed’ on the test and required to take it again. So, the question is how the assessment is being used. From a classroom perspective, it has been suggested that reporting total CPS competence profiles of the students may be more helpful to classroom teachers (e.g., Harding et al., 2017; Harding & Griffin, 2016). Nevertheless, considering multi-dimensional structures for CPS competence measures, could provide teachers with information about whether a student has an outstandingly different achievement on a particular skillset, and for some assessment purposes, this might be useful information for the teacher to have.

The small number of studies that examined evidence targeting the external aspect of validity focused mainly on self-reported constructs from student questionnaires such as perceived teamwork skills, performance in other subjects and students' teamwork skills as perceived by their teachers (e.g., Andrews-Todd & Forsyth, 2020; Rosen, 2015; Stadler, Herborn, et al., 2020a; Sun et al., 2020). For example, Stadler et al. (2020a) investigated the validity of the PISA CPS assessment tasks by identifying the extent to which a CPS competence measure is related to other collaboration measures (e.g., teacher-rated collaboration) and found moderate relations. However, to what extent CPS measures derived from PISA CPS tasks resemble students' real skills in solving problems in collaboration with others, as exhibited in interactions with humans, remains an open question.

In the present article, the CPS competence measures were only very weakly correlated with students' attitudes towards collaboration, and not enough evidence was found to support that those theoretically relevant constructs were indeed related. In another study, Andrews-Todd and Forsyth (2020) found that students' self-report of their collaborative preferences were not significantly correlated with their CPS skill profiles. Using PISA 2015 CPS data for four provinces in China, Tang et al. (2021) found that students who reported valuing relationships showed better CPS performance, while students who reported valuing teamwork tended to show worse CPS performance. This relationship was also evident in Models 2a-d of this paper's analysis, however, when taking student performance in the subject of science into account (Models 3a-d), this relationship was almost minimised. In their analysis, Tang et al. (2021) have not included any other student performance variables to explore the association with CPS performance. They concluded that students who valued interpersonal relationships would be highly motivated and would utilise more communication skills to ensure the collaboration to be carried out more smoothly, hence explaining the higher CPS performance (Tang et al., 2021). A potential explanation for the negative relationship between valuing teamwork and CPS performance has not been provided though.

CPS competence measures (overall and sub-scales) were found to be highly correlated with performance in other subject domains (i.e., science, mathematics, and reading), which might suggest that the cognitive aspect is more prevalent than the collaborative aspect in the social constructs developed. This is perhaps not surprising, since students were basically working on their own to complete the computer-simulated assessment, and it is possible that they were using their cognitive ability to complete it as expected to score high points. Hence, high performing students in computer-simulated assessments of subjects such as science, can be assumed to be also good enough to grasp what responses are needed in a computer-simulated problem-solving assessment that is independent of specific subject knowledge, without much recourse to social collaborative competences or dispositions. On the contrary, it might be argued that individual students' competence in, for instance, science is the result of successful learning through collaboration with others, and to this extent a good indicator of CPS, or at least as good as their self-reported attitudes to collaboration in a questionnaire context.

6.6.1 Implications for policy, practice, and research

There are several results with clear implications for educational policy, practice, and research.

This exploration shed light into the multidimensional structure of CPS competence, and it is concluded that evidence for CPS competence measures' structural validity aspect is important to be examined in future research, particularly after the inclusion of more items targeting all skills described as important for CPS competence. Results also showed that students' CPS competence measures (overall and sub-scales) have very weak correlation with students' attitudes towards collaboration. Limitations posed by the assessment design features, such as multiple-choice response options and computer-simulated agents instead of real humans, might have constrained what is possible to be measured as evidence for CPS competence in the PISA assessment.

There is still a question about whether measures derived from such controlled tests have external validity regarding students' capabilities to work together with others effectively on problem solving in "reality". So far, the literature has been limited to examining differences in student responses when communicating with computer-simulated partners versus real humans, with the means of communication being still constrained and deviating from natural communication. It is therefore suggested that researchers should focus on more in-depth analysis of student responses in real situations.

In addition, it may be more informative and productive for researchers to use results from PISA-like CPS assessments as additional pieces of information alongside the evaluations of teachers, instead of solely relying on them to show what students can do. Teacher evaluations can take the form of observing students working together towards problem solving in real situations. Finally, comparative judgement can be used for assessing students' work as an alternative method to traditional scoring (Jones et al., 2015). This method is based on collective expert judgements of students' work rather than item-by-item scoring schemes, and has been previously found to be well suited to assessing difficult-to-define skills such as mathematical problem solving (Jones et al., 2015; Jones & Inglis, 2015). Future work could involve the implementation of comparative judgement in the context of CPS extending the existing work related to mathematical problem solving.

6.6.2 Limitations

It is important to consider these findings in light of the limitations of this paper. First, the attitudinal measures used in the analysis are based upon student self-reports, hence, they could be affected by reporting inaccuracies as well as social desirability bias. Another limitation relates to the use of cross-sectional data available. Limitations to the statistical modelling analysis include listwise exclusion of missing data, which could lead to bias. Furthermore, multilevel modelling was not employed so further research could include school random effects in the model.

Additionally, limitations from a conceptual standpoint include the issue of item confidentiality. The availability of the CPS items' content (or the absence of it) has posed certain challenges in the independent use of the PISA 2015 data. Specifically, only a small proportion of CPS assessment items (about 10%), was released by test constructors for public view, meaning that users of the data must rely upon the descriptions of the test instrument provided. Two points in relation to item confidentiality are important to be made here. First, since CPS competence was an innovative domain for PISA 2015, this implies that there is no expectation for the material to be used in the following cycles again. Hence, there is less call for secure items as compared to the recurring domains (i.e., reading, science and mathematics). Second, PISA 2015 dataset does not include information about the possible actions that students make. Therefore, it is not possible to examine what is happening during the task for students who have made an 'incorrect' action. As others have previously argued (e.g., De Boeck & Scalise, 2019), interpretative information for the sample, such as think-aloud protocols, are also not provided. Finally, recent critiques of the traditional conceptualisation of gender as a binary construct should be acknowledged, although it is currently challenging to consider the whole landscape of genders (intersected with class, race/ethnicity, and cultures) using the PISA dataset.

6.7 Conclusion

In conclusion, this paper shed some light into the validity of the PISA 2015 CPS competence measure by analysing the available secondary data, which has been largely unused to date. The Rasch measurement framework and a unified validity definition were used to examine validity evidence for student CPS competence measures, based on the multidimensional character of CPS competence. Items were examined for fit, targeting, and measurement invariance using both a unidimensional model, reflecting overall CPS competence as well as three sub-scales with sets of items reflecting each hypothesised dimension. The investigation of the associations between CPS competence measures and other relevant constructs contributes to a very recent debate about the external validity aspect of PISA-like CPS assessments. Future research could investigate whether CPS competence measures influence student performance outcomes to provide further evidence for consequential validity.

The fact that such controlled assessment tools deviate from naturalistic, ecologically valid activities, raise questions about external validity and their ability to capture real collaboration processes. These are reasons for caution when using and interpreting PISA CPS competence measures. Finally, it follows that educational policy should be more sensitive to external and consequential validity considerations. Specifically, policy that typically focuses on results from large-scale international surveys to inform curriculum changes and educational reform regarding students’ CPS competence should consider the consequences of using such standardised CPS assessments, following their limitations.

6.8 References

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

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

Andrews-Todd, J., & Forsyth, C. M. (2020). Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Computers in Human Behavior*, *104*, 105759. https://doi.org/10.1016/j.chb.2018.10.025

Avvisati, F., & Keslair, F. (2014). *REPEST: Stata module to run estimations with weighted replicate samples and plausible values* (revised 05 Jun 2019) [Computer software]. Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s457918.html

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Lawrence Erlbaum.

Care, E., Griffin, P., & Wilson, M. (2018). *Assessment and Teaching of 21st Century Skills: Research and Applications* (1st edition..). Springer International Publishing. https://doi.org/10.1007/978-3-319-65368-6

Cukurova, M., Luckin, R., Millán, E., & Mavrikis, M. (2018). The NISPI framework: Analysing collaborative problem-solving from students' physical interactions. *Computers & Education*, *116*, 93–109. https://doi.org/10.1016/j.compedu.2017.08.007

De Boeck, P., & Scalise, K. (2019). Collaborative Problem Solving: Processing Actions, Time, and Performance. *Frontiers in Psychology*, *10*, 1280. https://doi.org/10.3389/fpsyg.2019.01280

Gao, Q., Zhang, S., Cai, Z., Liu, K., Hui, N., & Tong, M. (2022). Understanding student teachers' collaborative problem solving competency: Insights from process data and multidimensional item response theory. *Thinking Skills and Creativity*, *45*, 101097. https://doi.org/10.1016/j.tsc.2022.101097

Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 319–330. https://doi.org/10.1080/0969594042000304618

Graesser, A. C., Cai, Z., Morgan, B., & Wang, L. (2017). Assessment with computer agents that engage in conversational dialogues and trialogues with learners. *Computers in Human Behavior*, *76*, 607–616. https://doi.org/10.1016/j.chb.2017.03.041

Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the Science of Collaborative Problem Solving. *Psychological Science in the Public Interest*, *19*(2), 59–92. https://doi.org/10.1177/1529100618808244

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

Harding, S.-M. E., & Griffin, P. (2016). Rasch Measurement of Collaborative Problem Solving in an Online Environment. *Journal of Applied Measurement*, *1*(17), 35–53.

Harding, S.-M. E., Griffin, P., Awwal, N., Alom, B. M., & Scoular, C. (2017). Measuring Collaborative Problem Solving Using Mathematics-Based Tasks. *AERA Open*, *3*(3), 1–19. https://doi.org/10.1177/2332858417728046

Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? *Computers in Human Behavior*, *104*, 105624. https://doi.org/10.1016/j.chb.2018.07.035

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A Framework for Teachable Collaborative Problem Solving Skills. In P. Griffin & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 37–56). Springer Netherlands. https://doi.org/10.1007/978-94-017-9395-7_2

Jerrim, J., Oliver, M., & Sims, S. (2019). The relationship between inquiry-based teaching and students' achievement. New evidence from a longitudinal PISA study in England. *Learning and Instruction*, *61*, 35–44. https://doi.org/10.1016/j.learninstruc.2018.12.004

Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: Can comparative judgement help? *Educational Studies in Mathematics*, *89*(3), 337–355. https://doi.org/10.1007/s10649-015-9607-1

Jones, I., Swan, M., & Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*, *13*(1), 151–177. https://doi.org/10.1007/s10763-013-9497-6

Krkovic, K., Wüstenberg, S., & Greiff, S. (2016). Assessing Collaborative Behavior in Students: An Experiment-Based Assessment Approach. *European Journal of Psychological Assessment*, *32*(1), 52–60. https://doi.org/10.1027/1015-5759/a000329

Kuo, B.-C., Liao, C.-H., Pai, K.-C., Shih, S.-C., Li, C.-H., & Mok, M. M. C. (2020). Computer-based collaborative problem-solving assessment in Taiwan. *Educational Psychology*, *40*(9), 1164–1185. https://doi.org/10.1080/01443410.2018.1549317

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

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*.

Linacre, J. M. (2006a). *Dimensionality: Contrasts & variances*. https://www.winsteps.com/winman/principalcomponents.htm

Linacre, J. M. (2006b). *Reliability and separation of measures*. https://www.winsteps.com/winman/reliability.htm

Linacre, J. M. (2006c). *WINSTEPS Rasch measurement software*. Mesa Press.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.

Messick, S. (1995). Validity of Psychological Assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.

National Research Council. (2011). *Assessing 21st Century Skills: Summary of a Workshop*. National Academies Press (US). http://www.ncbi.nlm.nih.gov/books/NBK84218/

Nouri, J., Åkerfeldt, A., Fors, U., & Selander, S. (2017). Assessing Collaborative Problem Solving Skills in Technology-Enhanced Learning Environments – The PISA Framework and Modes of Communication. *International Journal of Emerging Technologies in Learning (IJET)*, *12*(04), 163. https://doi.org/10.3991/ijet.v12i04.6737

OECD. (n.d.). *Description of the Released Unit from the 2015 PISA Collaborative Problem-Solving Assessment, Collaborative Problem-Solving Skills, and Proficiency Levels*. Retrieved 29 January 2022, from https://www.oecd.org/pisa/test/CPS-Xandar-scoring-guide.pdf

OECD. (2009). *PISA Data Analysis Manual: SPSS, Second Edition*. OECD. https://doi.org/10.1787/9789264056275-en

OECD. (2017a). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. OECD Publishing. https://doi.org/10.1787/9789264281820-en

OECD. (2017b). *PISA 2015 Results (Volume V): Collaborative problem solving*. OECD Publishing. https://doi.org/10.1787/9789264285521-en

OECD. (2017c). *PISA 2015 Technical Report*. OECD Publishing. https://doi.org/10.1787/9789264273856-19-en

Oliveri, M. E., Lawless, R., & Molloy, H. (2017). A Literature Review on Collaborative Problem Solving for College and Workforce Readiness. *ETS Research Report Series*, *2017*(1), 1–27. https://doi.org/10.1002/ets2.12133

O’Neil, H. F., Chuang, S.-H. (sabrina), & Chung, G. K. W. K. (2003). Issues in the Computer-based Assessment of Collaborative Problem Solving. *Assessment in Education: Principles, Policy & Practice*, *10*(3), 361–373. https://doi.org/10.1080/0969594032000148190

Pampaka, M. (2021). Establishing Measurement Invariance across Time within an Accelerated Longitudinal Design. In A. Cernat & J. W. Sakshaug (Eds.), *Measurement Error in Longitudinal Data* (pp. 405–446). Oxford University Press. https://doi.org/10.1093/oso/9780198859987.003.0017

Pampaka, M., Williams, J., Hutcheson, G., Black, L., Davis, P., Hernandez-Martinez, P., & Wake, G. (2013). Measuring Alternative Learning Outcomes: Dispositions to study in Higher Education. *Journal of Applied Measurement*, *14*(2), 197–218.

Pólya, G. (1945). *How to solve it: A new aspect of mathematical method*. Princeton University Press.

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

Rosen, Y. (2015). Computer-based Assessment of Collaborative Problem Solving: Exploring the Feasibility of Human-to-Agent Approach. *International Journal of Artificial Intelligence in Education*, *25*(3), 380–406. https://doi.org/10.1007/s40593-015-0042-3

Rosen, Y., & Foltz, P. (2014). Assessing collaborative problem solving through automated technologies. *Research and Practice in Technology Enhanced Learning*, *9*, 389–410.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Scoular, C., & Care, E. (2020). Monitoring patterns of social and cognitive student behaviors in online collaborative problem solving assessments. *Computers in Human Behavior*, 105874. https://doi.org/10.1016/j.chb.2019.01.007

Scoular, C., Care, E., & Hesse, F. W. (2017). Designs for Operationalizing Collaborative Problem Solving for Automated Assessment. *Journal of Educational Measurement*, *54*(1), 12–35. https://doi.org/10.1111/jedm.12130

Scoular, C., Eleftheriadou, S., Ramalingam, D., & Cloney, D. (2020). Comparative analysis of student performance in collaborative problem solving: What does it tell us? *Australian Journal of Education*. https://doi.org/10.1177/0004944120957390

Smith, E. V. (2002). Understanding Rasch Measurement: Detecting and Evaluating the Impact of Multidimensionality Using Item Fit Statistics and Principal Components Analysis of Residuals. *Journal of Applied Measurement*, *3*(2), 205–231.

Stadler, M., Herborn, K., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: An investigation of the validity of the PISA 2015 CPS tasks. *Computers & Education*, *157*, 103964. https://doi.org/10.1016/j.compedu.2020.103964

StataCorp. (2019). *Stata Statistical Software: Release 16*. StataCorp LLC.

Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education*, *143*, 103672. https://doi.org/10.1016/j.compedu.2019.103672

Tang, P., Liu, H., & Wen, H. (2021). Factors Predicting Collaborative Problem Solving: Based on the Data From PISA 2015. *Frontiers in Education*, *6*. https://www.frontiersin.org/articles/10.3389/feduc.2021.619450

Webb, M., & Gibson, D. (2015). Technology enhanced assessment in complex collaborative settings. *Education and Information Technologies*, *20*(4), 675–695. https://doi.org/10.1007/s10639-015-9413-5

Wolfe, E. W., & Smith, E. V. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I - instrument development tools. *Journal of Applied Measurement*, *8*(1), 97–123.

Wolfe, E. W., & Smith, E. V. (2007b). Instrument development tools and activities for measure validation using rasch models: Part II - Validation activities. *Journal of Applied Measurement*, *8*(2), 204–234.

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

Wright, B. D., & Stone, M. (1999). *Measurement Essentials* (2nd ed.). Wide Range, Inc.

Yuan, J., Xiao, Y., & Liu, H. (2019). Assessment of Collaborative Problem Solving Based on Process Stream Data: A New Paradigm for Extracting Indicators and Modeling Dyad Data. *Frontiers in Psychology*, *10*, 369. https://doi.org/10.3389/fpsyg.2019.00369

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, *2012*(1), i–30. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x

Chapter 7 Validity of the PISA 2015 collaborative problem-solving assessment based on student response processes

7.1 Abstract

This paper uses the method of cognitive interviewing among students in England to address the lack of validity evidence of student response processes for the PISA 2015 collaborative problem solving (CPS) items. Currently, there have been no studies investigating the validity of the newly developed PISA 2015 CPS assessment tasks/test items using cognitive interviewing. This will be the first such study, using the task made publicly available in the PISA 2015 CPS assessment. Ten students from a secondary school in England completed the task and explained their responses. Grounded theory method was employed for data analysis, and five categories of student response processes were developed: ‘identifying contextual concerns’, ‘showing emotional intelligence’, ‘communicating in real life’, ‘providing alternative responses’, and ‘showing test savviness’. Results pointed to several weaknesses in the instrument, the PISA methodology and reporting, and its external and consequential validity. Students suggested their responses were not authentic in the sense of how they would respond in the 'real' situation being simulated. Instead, they were mediated by the simulation, and more specifically, by a rationality that they understood the system to demand. This contradicted the affective and affiliative tone that they would prefer if the computer-simulated partners were real, embodied co-participants in the problem-solving situation.

Keywords: collaborative problem solving, cognitive interviewing, validation, assessment, response processes, PISA

7.2 Introduction

Driven by the perceived needs of policy, student CPS competence assessment was introduced in the Programme for International Student Assessment (PISA) study in 2015. This study aimed to measure, and consequently ensure that students are equipped with, skills to meet the CPS demands of their future careers (OECD, 2017a). For the purposes of PISA 2015 CPS assessment, CPS competence was defined as “the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution” (OECD, 2017a, p. 134). Students’ CPS skills were assessed using an individualised computer-simulated assessment focusing on students’ performance in a collaborative event, as opposed to group performance. The main rationale for the PISA 2015 CPS assessment, was the need for a standardised summative assessment system, designed to provide large-scale information to countries about their student populations’ achievements.

The validity of the PISA 2015 CPS competence measure has been repeatedly questioned, mainly due to the constraints in communication imposed in the computer-based assessment. Specifically, to what extent do CPS competence measures derived from PISA 2015 CPS tasks resemble students’ real skills in solving problems in collaboration with others, as exhibited in interactions with humans? This remains an open question. Therefore, it is considered important to gather validity evidence and systematically investigate how students engage with the PISA 2015 CPS construct. A limited range of information is released by the PISA/OECD, e.g., items for only one out of the six PISA 2015 CPS tasks have been so far released, and descriptions of the possible student actions in the other CPS tasks are not available in the data set. Therefore, researchers do not have the information to examine what is happening during the CPS task completion process (De Boeck & Scalise, 2019).

Despite substantive research reporting the assessment of students’ CPS competence using computer-simulated assessment tasks (e.g., Graesser et al., 2018; Oliveri et al., 2017; Scoular et al., 2017; Sun et al., 2020), there has been only very limited exploration of evidence for their substantive, external, and consequential validity. As a method for gathering validity evidence, cognitive interviewing (CI) has been previously proved useful in identifying issues with existing PISA questionnaire items (Pepper et al., 2018). The evidence gathered during cognitive interviews can be used to check that the respondents interpret the items as intended by the item developer (Peterson et al., 2017). In a recent systematic review (CPS measurement review, Chapter 5), only one study was found to use cognitive interviewing to investigate the validity of a CPS competence measure (Siddiq & Scherer, 2017), and to the best of my knowledge, there has been no study exploring the validity of the PISA 2015 CPS competence measure through cognitive interviews to date.

The overall aim of this paper is to more deeply investigate issues related to validity evidence based on student response processes to one PISA 2015 CPS assessment task. In particular, I draw on cognitive interviews using verbal probing (Willis, 2005) with students interacting with the CPS competence construct. This method, although highly relevant for the validation of CPS competence measures, has been surprisingly neglected in the literature so far.

The paper proceeds as follows. The next section presents the background of the paper including the research questions, followed by an overview of methods. Results are then reported, followed by a discussion and conclusion.

7.3 Background

7.3.1 Validity evidence for computer-based CPS assessments

Due to PISA's far-reaching political influence, the investigation of the measurement tools used, and the validity of the data collected, is of particular importance. In a recent systematic review of PISA-related studies, Hopfenbeck et al. (2018) found that, despite their recognised value amongst researchers, PISA questionnaires have been criticised for various reasons.

Establishing validity and the interpretation of test scores entails taking into consideration multiple sources of relevant information (Karabenick et al., 2007). Messick's (1989, 1995) approach subsumes different validity categories within a comprehensive conception of validity that "is based on an integration of any evidence that bears on the interpretation or meaning of the test scores" (1995, p. 742). According to Messick, this includes content, substantive, structural, generalisability, external, and consequential sources of validity evidence. Most of the previous educational assessment research about the validity of CPS competence measures, which derived from computer-based assessments such as the PISA 2015 CPS assessment, is based on internal validity investigations, e.g., targeting generalisability, whereas aspects of validity such as structural, external, and consequential are less targeted (CPS measurement review, Chapter 5). Currently, there are no studies which systematically assess validity evidence based on student response processes in PISA's 2015 CPS assessment.

A recent systematic review (CPS concepts review, Chapter 4) showed that several CPS frameworks have been formed to guide the operationalisation of CPS competence and assessment development (e.g., Andrews-Todd & Forsyth, 2020; Hesse et al., 2015; OECD, 2017a; Sun et al., 2020). However, there is a limited number of studies about how the frameworks work in real life situations and how external factors might influence the results derived from the CPS assessments, which highlights the need for more empirical investigations (Nouri et al., 2017). For instance, Nouri et al. (2017) compared audio and text chats regarding human-to-human interactions and showed that some skills, as defined using the PISA 2015 CPS framework, were more visible with audio chat, which resulted in questioning the validity of the framework.

Additionally, a second systematic review (CPS measurement review, Chapter 5) showed that most of the existing CPS assessments using computer-simulated scenario-based tasks, were found to measure a limited spectrum of CPS skills, when mapped on existing CPS frameworks to investigate construct representation and targeting. Specifically, skills such as active listening, audience awareness, team learning, and team empowerment, were only scarcely or not at all covered. In another study, Scoular et al. (2020) compared three computer-based CPS assessments (including PISA 2015) to investigate the extent that they measure the same construct, and which aspects of CPS competence they target. It was found that skills related to negotiation and audience awareness were not well represented across three assessments of students' CPS competence. Such findings suggest that test developers have had limited success in eliciting these skills in computer-based assessment contexts to date.

Other limitations concerning the PISA 2015 CPS assessment have been recently highlighted (e.g., Graesser et al., 2018; Scoular et al., 2017). Specifically, the fact that participating students had to interact with computer-simulated partners rather than other students raised the concern of an assessment environment that deviates from naturalistic, ecologically valid CPS activities (Graesser et al., 2018). In addition, Scoular et al. (2017) argue that there are major limitations in the degree to which the interaction with computer-simulated partners can be regarded as capable of capturing communication competencies, and by extension capable of capturing the other social skills required for CPS that rely on communication. Given that communication is identified as having a central role in CPS, it is questionable whether or not a human-to-agent scenario would be able to elicit sufficient behaviours for good representation of the construct (Scoular et al., 2017).

Taking into account that the PISA 2015 CPS assessment scenarios constrain the number of possible discourse patterns (e.g., negotiation) to only one message exchange, this adds to the limitations of the assessment. Herborn et al. (2020) used the PISA 2015 CPS tasks to validate the assessment by investigating the effects of replacing computer-simulated partners with students. Funded by the OECD, Herborn et al.'s (2020) study obtained the otherwise confidential PISA 2015 CPS tasks and concluded that there is no significant difference between the types of collaboration partner. However, the allowance of external effects that occur in real human-to-human interactions was very limited, as the predefined chat communication was retained. Therefore, the generalisability of the results about the nature of the human-to-computer-simulated partner approach in resembling students' real CPS skills exhibited in interactions with humans is limited.

From the aforementioned limitations of the PISA 2015 CPS competence assessment, it could be argued that CPS is highly contextual in nature, and therefore a question that arises is: “How is it possible to be measured in practice?”, and when it is being measured by applying standardised methods and traditional criteria for scale construction: “How valid is the construct being measured?”. These are some of the questions that help shaping the aim of this paper and the research questions as outlined below.

7.3.2 The PISA 2015 collaborative problem-solving framework

The PISA 2015 CPS framework has been used to define and operationalise CPS competence, as a set of three newly conceptualised collaborative competencies: i.e., Establishing and maintaining shared understanding, Taking appropriate action to solve the problem, and Establishing and maintaining team organisation, and four problem-solving processes previously conceptualised in PISA 2012 (OECD, 2017a). “Establishing and maintaining shared understanding” relates to keeping track of what other team members know about the problem, their perspectives, and a shared vision of the problem states and activities. “Taking appropriate action to solve the problem” relates to performing actions, which can include physical actions and communication acts, that follow the appropriate steps to achieve a solution. “Establishing and maintaining team organisation” relates to helping to (re)organise the group by considering the knowledge, skills, abilities, and resources of group members, following the rules of engagement for roles in the group, as well as handling obstacles (OECD, 2017a).

These three competencies are crossed with the four individual problem-solving processes (i.e., exploring and understanding, representing and formulating, planning and executing, and monitoring and reflecting) to create a matrix of 12 cells, each representing a specific CPS skill (Table 7.1). Here, a ‘skill’ is a ‘collaborative competence’ utilised in a problem-solving ‘process’. Each item included in the CPS assessment is classified as targeting one of the CPS skills.

For the purposes of the PISA 2015 study (e.g., standardisation, cross-national comparisons), the assessment of students’ CPS competence was operationalised within a computer-based environment and CPS competence was assessed by evaluating how well the individual student collaborated with computer-simulated partners during the problem-solving process (OECD, 2017a). The assessment included a set of computer-based tasks which used computer-simulated partners to replace human group members and a selection of pre-defined written messages to replace open communication.

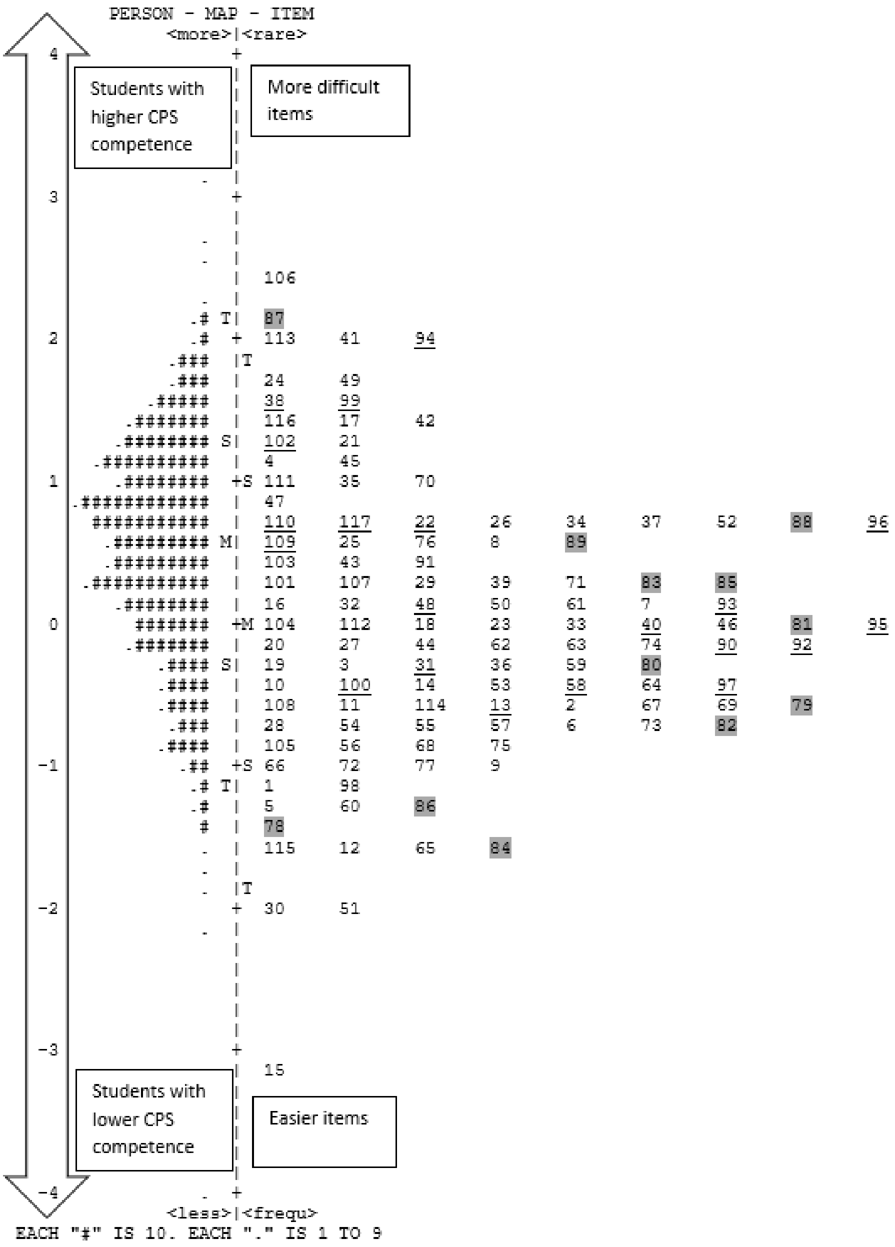
Table 7.1. PISA 2015 Collaborative problem-solving framework (OECD, 2017a)

	(1) Establishing and maintaining shared understanding	(2) Taking appropriate action to solve the problem	(3) Establishing and maintaining team organisation
(A) Exploring and understanding	(A1) Discovering perspectives and abilities of team members **20 items**	(A2) Discovering the type of collaborative interaction to solve the problem, along with goals **2 items**	(A3) Understanding roles to solve the problem **0 items**
(B) Representing and formulating	(B1) Building a shared representation and negotiating the meaning of the problem (common ground) **24 items**	(B2) Identifying and describing tasks to be completed **5 items**	(B3) Describing roles and team organisation (communication protocol/rules of engagement) **8 items**
(C) Planning and Executing	(C1) Communicating with team members about the actions to be/being performed **5 items**	(C2) Enacting plans **16 items**	(C3) Following rules of engagement (e.g., prompting other team members to perform their tasks) **14 items**
(D) Monitoring and reflecting	(D1) Monitoring and repairing the shared understanding **12 items**	(D2) Monitoring results of actions and evaluating success in solving the problem **3 items**	(D3) Monitoring, providing feedback and adapting the team organisation and roles **8 items**

Notes: Presented in bold are the number of items designed to assess each CPS skill in the PISA 2015 CPS assessment. Own adaptation of the framework based on information found in PISA’s technical report (OECD, 2017c).

7.3.3 Information about Xandar items from Chapter 6

In Chapter 6, CPS tasks included in the PISA 2015 CPS assessment have been used to construct a CPS competence measure for the analytical sample of students in England. Results for the Rasch analysis conducted in Chapter 6 can be used to understand the difficulty of the visible CPS (Xandar) items. Specifically, Figure 7.1 presents the person-item map of the overall CPS competence scale.



```
                 PERSON - MAP - ITEM
                     <more>|<rare>
 4                         +
                           |
                           |
                           |
                           |
                           |
                        .  |
 3                         +
                           |
                        .  |
                        .  |
                           |  106
                        .  |
                      .# T|  87
 2                    .#   +  113     41     94
                    .###   |T
                    .###   |  24      49
                  .#####   |  38      99
                .#######   |  116     17     42
               .######## S|  102     21
             .##########   |  4       45
 1             .########  +S  111     35     70
          .############   |  47
            ###########   |  110     117    22     26     34     37     52     88     96
             .######### M|  109     25     76     8      89
             .#########   |  103     43     91
           .###########   |  101     107    29     39     71     83     85
               .########   |  16      32     48     50     61     7      93
 0               #######  +M  104     112    18     23     33     40     46     81     95
                .#######   |  20      27     44     62     63     74     90     92
                   .#### S|  19      3      31     36     59     80
                   .####   |  10      100    14     53     58     64     97
                   .####   |  108     11     114    13     2      67     69     79
                    .###   |  28      54     55     57     6      73     82
                   .####   |  105     56     68     75
-1                   .##  +S  66      72     77     9
                      .# T|  1       98
                      .#   |  5       60     86
                       #   |  78
                        .  |  115     12     65     84
                        .  |
                        .  |T
-2                         +  30      51
                        .  |
                           |
                           |
                           |
                           |
                           |
-3                         +
                           |  15
                           |
                           |
                           |
                           |
                           |
-4                      .  +
                     <less>|<frequ>
 EACH "#" IS 10. EACH "." IS 1 TO 9
```

Figure 7.1. Person-item map for the overall CPS competence including 117 items

Notes: Items with underline have three or more response categories. Items highlighted with grey are the visible Xandar items.

Information from the analysis conducted in Chapter 6 can be used to inform the methodology of this chapter. In particular, the person-item map can inform about which Xandar items students in England that took the PISA 2015 CPS assessment found difficult and which they found easy to response correct to. As shown in Figure 7.1, the most difficult item of the task Xandar for this analytical sample is item 87, which is located towards the top of the item distribution. This item aims to assess students' CPS skill 'Monitoring and repairing the shared understanding'. To better understand why this is a difficult item, its content was reviewed. Specifically, students had to notice a change in the problem space, which violated the previously agreed rules of engagement. Its difficulty may be therefore explained because students are required to track not only the chat space but also the change in status in the problem space (OECD, 2017b). In addition, the credited response does not sound overtly collaborative (i.e., "I should answer the Geography questions. Let's work on the subjects we chose."), which makes it more difficult (OECD, 2017b).

As shown in the person-item map, the easiest item is item 84, which is located towards the bottom of the item hierarchy. This item aims to assess students' CPS skill 'Describe roles and team organisation' and requires students to evaluate the reasons provided by each team member for claiming a subject. Given that one of the team members has given a reason that would be an advantage for the team, this item is not overly difficult (OECD, 2017b). The credited response is also clearly collaborative as it asks for the other team member's agreement with the proposed approach (i.e., "It sounds as though People should be Alice's subject. Zach, are you OK with that?").

7.3.4 Cognitive interviewing and the PISA study

It has been previously argued that cognitive interviewing (CI) is a powerful method for understanding the thought processes of the respondents when answering questions (Beatty & Willis, 2007). The usefulness of CI depends not so much on the type of question to evaluate, as the processing that the respondent performs (Castillo-Díaz & Padilla, 2013). Regardless of the timing of its use, CI can provide valuable information about the cognitive operations underlying item interpretation and response and the consequent validity of test interpretation (Peterson et al., 2017). The respondent's interpretation is foundational to the inferences made from assessment results, and therefore, misinterpretation of test items directly affects test validity (American Educational Research Association et al., 2014).

A search of the literature across all disciplines using the academic database Scopus shows the growth in the use of CI. Of the total of 2,443 references identified (October 2021) across all disciplines including the term "cognitive interview" or "cognitive interviewing" in their abstract, title or keywords, about 40 per cent of them were associated with the subject area of Medicine, while almost 15 per cent were associated with the subject area of Social Sciences[27]. The majority of these 2,443 references were published from 2010 onwards, growing exponentially to a peak of 266 academic outputs in 2021 (see Appendix 21). To explore how, and to what extent, this method of data collection has been used in combination with the PISA study, or data derived from it, the literature search was constrained to studies also including the term "PISA" or "Programme for International Student Assessment" in their abstract, title or keywords. From this updated search, only 3 studies were found to include both terms (Benítez & Padilla, 2014; Hopfenbeck & Maul, 2011; Pepper et al., 2018). Interestingly, no study was found to use CI to validate the PISA 2015 CPS assessment scale.

The primary role of CI in these three studies was to validate different PISA scales. In the first study, Hopfenbeck and Maul (2011) used CI among students in Norway to address the lack of validity evidence of student response processes for the PISA 2006 self-regulated learning in science items. Findings from this study suggest that a non-trivial proportion of students were not responding to questionnaire items in the desired manner. Authors concluded with a note of caution concerning the interpretation of results from the PISA questionnaire scales under investigation. In the second study, Benítez and Padilla (2014) investigated differential item

[27] Scopus database includes 'Education' subject area under Social Sciences, and for that reason separate results are not presented for Education here.

functioning sources in PISA 2006 student questionnaire scales by using a mixed method design. Participants from Spain and the United States were recruited for CI, and the results indicated that inferences about the differences or similarities between groups in some items should be established with caution. Authors argued that CI has provided information that would not have been accessed by the implementation of the statistical methods alone. Finally, Pepper et al. (2018) used items from the PISA 2012 student self-efficacy in mathematics scale as the basis for CI investigation in three education systems (England, Estonia, and Hong Kong). CI was proved to be a useful method in identifying issues with translation and comprehension, and it was concluded that the OECD should conduct such validations of the PISA questionnaire items, ahead of their use in future PISA cycles.

7.3.5 Research questions

There appears to be lack of studies systematically examining student response processes to the PISA 2015 CPS competence assessment and the external aspect of validity concerning the derived measures. Therefore, there is a clear need for a more comprehensive validation of such computer-based CPS assessments. This paper aims to contribute to the validation of the PISA 2015 CPS competence assessment adopting a unified validity definition (Messick, 1995) and using cognitive interviewing (Willis, 2005) to gather validity evidence of response processes for the PISA 2015 CPS items.

The following research question and sub-questions guided this paper:

> RQ4. What does the PISA 2015 CPS assessment actually measure according to student perspectives?
>
> > RQ4.a: How do students comprehend the CPS assessment items and how do they explain their answers to them?
> >
> > RQ4.b: What are the implications for the external validity of the CPS assessment?

7.4 Methods

7.4.1 Participants

Cognitive interviews were conducted with 10 students (5 males and 5 females), aged between 15 years-old and 15 years and 7 months old, attending a secondary school in Greater Manchester, England. The participants were recruited via their mathematics teacher. The criteria that guided the selection of participants were: equal gender distribution; age range close to 15 years-old, similar to the target population of the PISA study[28]; balanced distribution of educational level based on students' grade in the Mathematics and English subjects; and variety of social behaviour in class as described by their teachers. All the participating students are referred to by pseudonyms (Table 7.2).

Table 7.2. Participant information

Student pseudonym	Gender	Interview duration
Anna	Female	25 minutes
Becky	Female	38 minutes
Ella	Female	22 minutes
Emily	Female	35 minutes
Maria	Female	28 minutes
John	Male	39 minutes
Leo	Male	23 minutes
Oliver	Male	36 minutes
Pablo	Male	31 minutes
Stephan	Male	40 minutes

Unlike psychometric methods used in establishing evidence of validity, varying perspectives rather than representativeness is the goal when sampling for cognitive interviews (Beatty & Willis, 2007; Willis, 2005). For that reason, students were selected with the aim to cover a variety of demographic characteristics, social behaviour, and prior attainment. Parents/guardians received a letter with

[28] PISA assesses students between the ages of 15 years and 3 months and 16 years and 2 months, and who are enrolled in an educational institution at grade 7 or higher (OECD, 2017).

information and gave consent to allow their children to be interviewed. All participating students read a letter with participant information and gave assent to be interviewed. The participants were guaranteed confidentiality and that the data would be used solely for purposes related to research.

7.4.2 Assessment task – PISA 2015 CPS assessment

Limited information is released by the PISA/OECD, i.e., items for only one out of the six CPS tasks included in the PISA 2015 CPS assessment. Due to this restriction, this paper makes use of the only publicly available PISA CPS task named 'Xandar'[29]. Each student completed the PISA 2015 CPS assessment on a computer individually. In the PISA CPS task Xandar, a three-person team consisting of the student test-taker and two computer-simulated partners (Alice and Zach) takes part in a contest where they must answer questions about the fictional country of Xandar. The questions are evenly divided between Xandar's geography, people, and economy. The task is consisted of 12 items each targeting one of the CPS skills from the PISA 2015 CPS framework. The introduction of the Xandar task informs students as follows (OECD, 2017b, p. 53):

> "Your teacher has divided the class into three-person teams for a contest. The winning team will be the first to correctly answer 12 questions about the country of Xandar. Answers can be found by opening links on a map of Xandar."

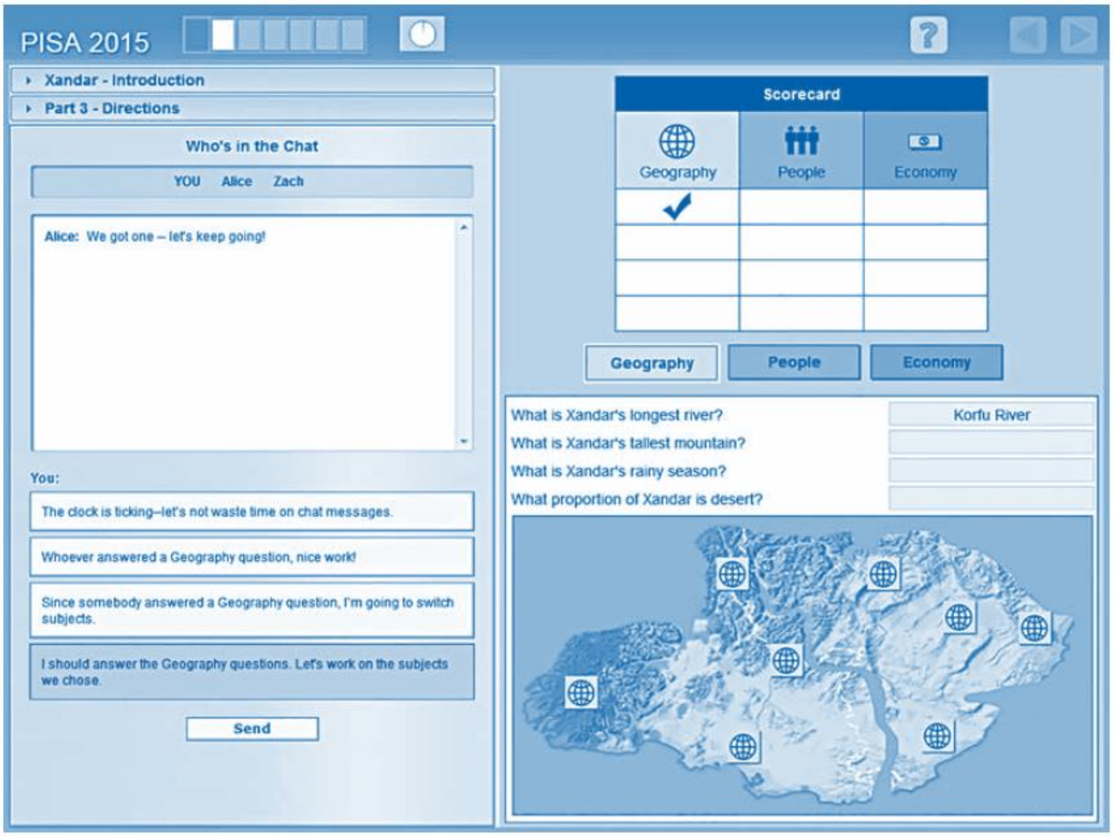
The Xandar task has the following four parts (OECD, 2017b):

> Part 1 – Agreeing on a strategy. In this part, the student is familiarised with how the contest will proceed, the chat interface and the task space (buttons that students can click and the scorecard that monitors team progress). The student has been assigned to work in a team with Alice and Zach and the

[29] https://www.oecd.org/pisa/test/other-languages/xandarurlreplacementtest.htm

> teacher has asked teams to put off searching for questions and answers until the contest begins and instead to discuss how to approach the contest.
>
> Part 2 – Reaching consensus regarding preferences. In this part, the student is informed that each group member will be responsible for the questions in one subject area. Alice and Zach begin by showing their preference for taking the same subject. The student is expected to help resolve this disagreement.
>
> Part 3 – Playing the game effectively. In this part, the student is informed that their assigned subject area is geography, regardless of whether they claimed it for themselves in the previous part. Before a student has a chance to try and answer a geography question, a computer-simulated partner violates the agreement and answers one of them. The student is required to track not only the chat but also the change in the status in the problem space.
>
> Part 4 – Assessing progress. In this part, the student is required to evaluate the team's progress and fix any problems that have resulted. Regardless of the student's answer, Zach indicates experiencing trouble answering questions in the assigned subject area and the student is required to present a proposal that is most effective in working towards the problem solution. Finally, regardless of how the student responded to the last item of the task, they are informed that their team won the contest by answering all the questions correctly and the unit ends.

To answer items in the Xandar task, students could either make a multiple-choice selection of predefined messages presented in the chat space or perform actions (e.g., dragging and dropping) in the task space (OECD, 2017b). No free-response items were available to students and items were independent of one another. Figure 7.2 illustrates item 87, which was previously found to be the most difficult item in the Xandar task for the PISA 2015 student sample for England (Chapter 6). In this item, a question in Geography (subject previously assigned to the student test-taker) is ticked automatically as correct before the student has a chance to try and answer. This item requires students to notice that the event in the problem space violates the previous agreement (OECD, 2017b). The credited response: “I should answer the Geography questions. Let’s work on the subjects we chose” is claimed to balance the problem-solving demands and the team’s assigned roles in the game (OECD, 2017b). For further information see the official OECD reports on PISA 2015 results (OECD, 2017b) and scoring guide for the released CPS task (OECD, n.d.).



Alice: We got one – let's keep going!

You:

The clock is ticking–let's not waste time on chat messages.

Whoever answered a Geography question, nice work!

Since somebody answered a Geography question, I'm going to switch subjects.

I should answer the Geography questions. Let's work on the subjects we chose.

What is Xandar's longest river?	Korfu River
What is Xandar's tallest mountain?	
What is Xandar's rainy season?	
What proportion of Xandar is desert?	

Figure 7.2. Screenshot of a released PISA 2015 CPS item (OECD, 2017b)

Notes: Chat space (left) displays the pre-defined messages for communication with the computer-simulated agents, and task space (right) is where actions are performed. Forth message (highlighted) is the credited response representing the CPS skill ‘Monitoring and repairing the shared understanding’. Material used under Creative Commons Attribution-NonCommercial-ShareAlike 3.0 IGO (CC BY-NC-SA 3.0 IGO) license.

7.4.3 Cognitive interviewing

Cognitive interviewing (CI) is the main approach for qualitative data collection in this paper. It is a method that obtains evidence on response processes and interpretations by respondents when answering survey questions/test items (Beatty & Willis, 2007; Willis, 2005). It can be used in scale development to inform item revision decisions and can provide evidence of validity based on test content and respondents' response processes (American Educational Research Association et al., 2014; Peterson et al., 2017). In CI, the interviewer asks respondents to describe their thinking either concurrently as they answer each question or retrospectively after they complete all questions. The goal is to identify items where there is a misalignment between participant interpretation and the developer's intentions and to identify ways to modify those items based on participants' responses (Peterson et al., 2017). Based mainly on a cognitive four-stage model, CI explores the various stages of the question-and-answer process, which are: comprehension, recall, judgment, and response (Tourangeau, 1984). The actual response selected by the respondent, in and of itself, is argued to provide no evidence concerning whether the respondent has engaged in the cognitive processes, while an invalid response could occur due to a breakdown at any stage of this process (Hopfenbeck & Maul, 2011). Additionally, a simple lack of motivation and perceived social desirability, are both reasons why students may not fully engage in the question-and-answer process.

One of the first steps involved in CI is the identification of item intent prior to the CI (Peterson et al., 2017). Table 7.3 presents a detailed account of the intent and scoring of the CPS items included in PISA 2015 CPS task. Item intent directly pertains to the aspect of the construct the item is designed to tap and forms the basis from which to judge if there is a misalignment between how the respondent interprets the item and what it is intended to measure (Peterson et al., 2017).

Table 7.3. Intent and scoring of items in the PISA 2015 CPS task

Item name	Item number	Credited item response option	Other item response options	Item's intent	CPS skill
X1	78	Click on the "Join the Chat" button.	Click on other active buttons on the task space.	Item requires student to respond to the directions on the screen.	(C3) Following rules of engagement
X2	79	*Maybe we should talk about strategy first.*	*I wonder if some of the other teams have started yet.* *I hope the questions are easy.* *Alice, you can see what to do once we get started.*	Item requires student to take the initiative to suggest the first logical step required to solve the problem.	(C1) Communicating with team members about the actions to be/being performed
X3	80	*True, but what's a good way to do that?*	*Right, the first team to answer all the questions wins.* *Do you think all the teams have to answer the same questions?* *First we should find out what we'll get for winning the contest.*	Item requires student to focus the discussion on how best to meet the goal of the contest and solicit ideas from the team.	(B1) Building a shared representation and negotiating the meaning of the problem (common ground)
X4	81	*We can answer more questions if we divide them among us.*	*The rules of the contest seem pretty simple. Let's just do our best.* *We can each work our fastest, but some of us will still be faster than others.* *It doesn't matter whether one of us answers more questions than the others, so long as we win.*	Item requires student to volunteer information not specifically requested by the other team members to help the team devise a strategy.	(B1) Building a shared representation and negotiating the meaning of the problem (common ground)

X5	82	*We could each take one of the subjects.*	*If there's a prize for winning, let's divide it equally.* *The contest lets us come up with our own team strategy.* *OK, then we're ready to begin.*	Item requires student to confirm and slightly extend the approach that has been agreed upon.	(B3) Describe roles and team organisation (communication protocol/rules of engagement)
X6	83	*Can each of you explain why you want that subject?*	*Nobody asked me what subject I want. Why should you guys choose first?* *Why are we wasting time arguing about this?* *Alice and Zach, are you going to answer questions faster than you choose subjects?*	Item requires student to help team members negotiate a solution when a conflict arises.	(A1) Discovering perspectives and abilities of team members
X7	84	*It sounds as though People should be Alice's subject. Zach, are you OK with that?*	*Alice, maybe you could study abroad in a visiting students program.* *Yes, it's good to know what your interests are.* *People in Xandar probably aren't very different from people anywhere else.*	Item requires student to evaluate the reasons provided by each team member.	(B3) Describe roles and team organisation (communication protocol/rules of engagement)
X8	85	*I'll take Geography.*	*Well, everyone likes money.* *Liking money doesn't mean you understand economy.* *We need to stop debating and make a decision.*	Item requires student to assume responsibility for identifying the one remaining subject area that needs to be claimed.	(B3) Describe roles and team organisation (communication protocol/rules of engagement)

X9	86	Click on the "Geography" button.	Click on other active buttons on the task space.	Item requires student to act based on the agreed-upon role, respond to directions on the screen, and click the correct button.	(C3) Following rules of engagement
X10	87	*I should answer the Geography questions. Let's work on the subjects we chose.*	*The clock is ticking - let's not waste time on chat messages.* *Whoever answered a Geography question, nice work!* *Since somebody answered a Geography question, I'm going to switch subjects.*	Item requires student to notice that the event in the problem space violates the agreement that each team member would take one of the subjects.	(D1) Monitoring and repairing the shared understanding
X11	88	*We look fine, except for Economy.*	*I think your scorecard is working – mine is.* *Great, we're half way there.* *I'm not sure since I don't know the other teams' scores.*	Item requires student to respond to a question from one team member and also provide additional information about how team is progressing.	(D2) Monitoring the results of actions and evaluating success in solving the problem
X12	89	*Keep trying. When Alice and I are done we'll help you – right Alice?*	*Zach, aren't you the one who said we all had to work fast?* *Do you expect us to stop what we're doing and help you instead?* *Are you behind because you were working on my Geography questions?*	Item requires student to present a proposal that is most effective in working towards the problem solution.	(D3) Monitoring, providing feedback and adapting the team organisation and roles

Notes. Table produced using information from OECD (2017b) and OECD (n.d.)

7.4.4 Verbal probing

Verbal probing is a core verbal reporting technique that involves the interviewer following up (either immediately or at the end of the interview) student's response to a target question by probing for other specific information relevant to the question or to the specific answer given (Willis, 2005). A key decision in verbal probing concerns the choice of when to probe and what probes to develop. For this paper, retrospective probing was used to create an environment that closely approximated the testing situation and eliminated the disruption in students' responses (Beatty & Willis, 2007). Prior to the interview, items were appraised considering the four cognitive processes, and a list of anticipated probes was developed to target potential areas of confusion. An example anticipated probe for the Xandar item presented earlier is "Was there an answer you wanted to give that was not available in the response options?". For more details see Appendix 8. Spontaneous probes were also used as the interviews progressed.

7.4.5 Procedure – Interview protocol

Cognitive interviews were conducted on a one-to-one and face-to-face basis in a quiet room at the participant's school setting within normal school hours. The fieldwork was undertaken between June 2019 and July 2019. For the total number of cognitive interviews conducted, I was the only interviewer. The interviews lasted from approximately 22 minutes to 40 minutes with the time split between the time students needed to complete the task and the time allowed for verbal probing. Students completed the Xandar task on a laptop provided by the University of Manchester with their responses being screen recorded. The captured screen-recording video was used for reflection during verbal probing.

The interview protocol consisted of five steps (see Chapter 3 for a detailed description). At the beginning of each interview a brief introduction of the purpose of this research was provided. In the second step, students were asked to complete the CPS task on the laptop. Once students reached the end of the task I moved to the third step, which involved asking them to explain how they understood every item and how they have gone about answering them.

This step was structured to cover two main verbal probes, which were asked to all students for every item:

1) How do you understand this part/statement?
2) Can you explain why you have given this answer?

A follow-up to the first verbal probe was "Can you tell me a little more about what is happening in this part?" And a follow-up to the second was "Can you tell me a little more about why you chose that answer?" Depending on their response, students were then asked some anticipated probes from the list that has been developed in advance (see Appendix 8) as well as spontaneous probes such as "What do you mean?". The fourth step included open questions about students' experience with the assessment in general. At the close of the interviews (fifth step), students were given the chance to ask questions about their interview or add any last thoughts. The interview protocol used to conduct all interviews in this paper is detailed in Appendix 11.

7.4.6 Data analysis

There is currently no standard method of analysis for CI (Peterson et al., 2017). Analytic techniques range from less intensive, e.g., making notes as the respondent is speaking, to more detailed coding schemes (Willis, 2005). In this paper, grounded theory and more specifically constant comparative methods are employed for the analysis of student response processes to CPS items. Grounded theory "uses a systematic set of procedures to develop and inductively derive grounded theory about a phenomenon" (Strauss & Corbin, 1998, p. 24). It consists of at least two main phases: an initial phase involving naming each word, line, or segment of data followed by a focused, selective phase that uses the most significant or frequent initial codes to sort, synthesise, integrate, and organise large amounts of data (Charmaz, 2006). Constant comparison involves taking information from data collection and comparing it to emerging categories during each stage of analysis to find similarities and differences (Creswell, 1998). The interview data was analysed through the following stages: familiarisation, reflection, initial coding, focused coding, and axial coding. Theoretical sensitivity was a critical part of all the coding stages. When coding, I kept thinking about actions and processes, not of individuals, as a strategy in constructing theory and moving beyond categorising types of individuals (Charmaz, 2006). In addition, I built a strong theoretical framework by reviewing how CPS has been conceptualised in the literature and by reviewing extended literature in the field of CPS more generally. The software NVivo 12 (QSR International Pty Ltd., 2018) was used to assist with coding.

7.4.6.1 Familiarisation

Familiarisation with interview data started by reading and rereading the interview transcripts and notes. At this stage, "interesting response examples" or tentative broad codes were briefly noted for each interview. These written records served as reminders of what I have captured during the reading process.

7.4.6.2 Reflection

At the stage of reflection, some preliminary cross-case analyses were conducted. This process was carried out by comparing and critically evaluating individual case data with other cases. Corbin and Strauss (1990) state that making comparisons can assist the researcher in guarding against bias and help to achieve greater precision and consistency. During this process, some important questions were asked, such as "Do the ideas in this case differ from other cases?", and "Are there any new ideas emerging from the case data?". By asking and answering these questions, similarities and differences among cases were highlighted, and were recorded in memos.

7.4.6.3 Initial coding

Initial coding helps to separate data into categories and to see processes (Charmaz, 2006). During this phase, the aim was to remain open to what the material suggests and make codes fit the data rather than forcing the data to fit the codes (Charmaz, 2006). At the same time, a set of concepts from the PISA 2015 CPS theoretical framework, which were identified when reviewing items' intent, were also used as potential codes. Overall, initial codes remained provisional, comparative, and grounded in the data (Charmaz, 2006). A total of 80 initial codes were developed during the initial coding phase, with various references/quotes assigned to each one of them, ranging from one to 40 references (for more details see Appendix 12). An initial coding example is presented below with initial codes in bold and quotes helping to formulate the codes being underlined.

> Interview excerpt: *You have to understand, you have to <u>listen to their reasons</u> for why they wanted People [**evaluating reasons**] and <u>allocate which one to each,</u> which one to who you think fits best [**assigning roles**]. From their answers I thought Alice, because she had <u>a genuine reason</u> and a passion for it [**evaluating reasons**].* (John)

7.4.6.4 Focused coding

Focused coding was the second major phase in coding, after the initial coding, which aimed to help synthesise and explain larger segments of data. Focused coding is more directed, selective, and conceptual than initial coding (Glaser, 1978). It involves using the most significant and/or frequent earlier codes to shift through large amounts of data (Charmaz, 2006). One goal is to determine the adequacy of those codes. It also requires decisions about which initial codes make the most analytic sense to categorise the data (Charmaz, 2006). Initial codes that were only assigned to one reference were likely candidates to be grouped with other relevant codes. A total of 13 focused codes were developed through comparing data to initial codes and across interviews. For example, codes such as 'answering faster' and 'saving time' were grouped together under the focused code 'getting the task done quicker' and reflected the fact that students considered the time limits of the competition to give a response. Appendices 13 and 14 present how the initial codes have been grouped to develop the focused codes and the full list of focused codes.

7.4.6.5 Axial coding

Axial coding is Strauss and Corbin's (1998) strategy for bringing data back together again in a coherent whole. Axial coding follows the development of a major category, although it may be in an early stage of development (Charmaz, 2006). While engaged in axial coding, Strauss and Corbin (1998) apply a set of scientific terms, such as conditions, actions/interactions, and consequences, to group participants' statements into components of an organising scheme. The categories that were developed are the following: identifying contextual concerns, showing emotional intelligence, communicating in real life, providing alternative responses, and showing test savviness, which are presented in detail in the following Results section.

7.5 Results

This section presents the categories derived from the analysis of student cognitive interviews reflecting how they understood and responded to the CPS items. Table 7.4 provides an overview of the five categories, which are detailed next.

Table 7.4. Summary of the five categories of student response processes

Category	Description	Focused codes
1) Identifying contextual concerns	Based on contextual information offered in the task, students find non-credited responses to be better fit for purpose.	approaching the task, evaluating reasons, understanding teammates
2) Showing emotional intelligence	Students consider others' feelings when responding and act towards retaining a balance and good team spirit in the group.	encouraging team, evaluating response options, understanding the tone of responses, working in or as a team
3) Communicating in real life	Students reflect on their responses if the scenario was taking place in real life.	evaluating response options, feeling uncomfortable, understanding the tone of responses
4) Providing alternative responses	Students volunteer responses outside the list of messages provided in the task.	approaching the task, evaluating response options, feeling uncomfortable, proposing plan, providing clarification
5) Showing test savviness	Students being able to exclude obvious responses that have no credit.	approaching the task, evaluating response options, getting the task done quicker, understanding the tone of responses, working in or as a team

7.5.1 Identifying contextual concerns

In item X8 the student has identified contextual concerns that seem to be valid, whereas the assessment is focused on the 'generic' processes. Specifically, in this item Zach claims the Economy subject by saying "I guess Economy would be all right. I like money.", and the student is expected to take responsibility for the last subject (Geography) by selecting the credited response "I'll take Geography". However, the credited response does not account for a reflection of Zach's suitability for the subject that he claimed, as was the case in the previous item (item X7), in which the student was expected to evaluate the reasons that the two teammates gave for claiming the subject People and assign it to Alice as the most well-suited one. Interestingly, both items target the same CPS skill "Describing roles and team organisation". The following examples go to the heart of the problem of validity. Instead of the credited response, students selected the message "Liking money doesn't mean you understand economy.", since they were not persuaded by the reasons given by their teammate to claim that subject.

> *Zach said that economy would be ok, he doesn't give any reasons it's just that he likes money. I would have been more confident saying "I'll take Geography" if he would have given more reasons as to why he would be happier with economy. Because he is not really giving me any confidence that he would be good at this subject, so I thought I'd say that [response: "Liking money doesn't mean you understand economy"].* (Emily)

> *He wants the Economy subject because he likes money, and he doesn't really take into consideration about the Geography one and if he might know anything about that. And he doesn't say that he knows anything about the Economy one, he just chooses it because he likes it.* (Becky)

7.5.2 Showing emotional intelligence

Student responses showed how it is important to take the feelings into account when deciding about their response. For example, item X11 required students to monitor the progress that the team has made in answering the questions about Xandar. As shown in the scorecard that was displayed in the problem space, half of the questions were answered correctly. However, the correct answers were given only in the subjects of People (assigned to Alice) and Geography (assigned to the student/test taker), while Economy (assigned to Zach) had no correct answers. Alice asked "How are we doing?" in the chat and the credited response was "We look fine, except for Economy." The item tests providing information about the team's progress as well as identifying the area that lacks progress. Two students, who have selected the credited response, have also demonstrated emotional intelligence.

> *I was honest so we could get the best outcome. […] It's not very nice, it's a bit like making the other person feel guilty.* (John)

> *If it was in real life, I would have felt quite harsh saying that [response: "We look fine, except for Economy."], but it was the truth, so I said that. Because Alice knew what she is talking about, because she said she likes the People subject and she researched about it, but Zach didn't know anything about the Economy one.* (Becky)

The term "real life" signals that there is in reality a lot more to be considered than the "game" here. The way the participants chose to respond in the assessment situation could be different from what they would have said to their friends, who would not take offense by that response, compared to some other team member who they do not know. This emotional intelligence demonstrated in students' explanations is important to be considered as part of validity evidence. Although students selected the credited response that addressed the area lacking progress, they recognised that sending that message, could potentially hurt the feelings of the person working on that area. This suggests that others might opt for the same answer without this empathy, while others might be put off this answer because of a sense that this is not the right empathic thing to do. This is demonstrated in the following two contrasting examples.

> *Zach hasn't done anything, that's what it looks like, but me and Alice have been doing the work basically. I didn't want not to address that he hasn't done anything. (Anna)*

> *It seems like the Geography area is going well and so is the People one, but the Economy is lacking because there is not ticks in it [the scorecard]. I was originally going to say the third one [response: "We look fine, except for Economy"], but I feel like that can seem like I am attacking. Even though it's evidently lacking, I still feel like there is no point in putting that person down, because that's just going to make it harder, because instead of giving them confidence, you sort of knocking it. So, it's better to boost people's confidence instead of making them feel worse.* (Emily)

In item X10, a checkmark is placed on the scoreboard to indicate that one of the questions on Xandar's Geography has been answered (this was done automatically by the assessment system and not because of the student/test taker's actions) and Alice sends a message saying that the team got one question right in Geography. Students must then come up with an appropriate response. The item tests whether the student has observed that the previously agreed rules of engagement, i.e., that the student himself or herself should answer the questions related to Geography, are not being followed. Therefore, students are expected to realise that the plan has been violated by one of the teammates and to act towards repairing the shared understanding by reminding the teammates about the plan. For that reason, the credited response was "I should answer the Geography questions. Let's work on the subjects we chose." However, students who noticed the violation in the agreement said that they avoided selecting the 'credited' response as it would create arguments in the group. Instead, they preferred the message "Whoever answered a Geography question, nice work!" to praise their teammates for the correct answer and the progress made.

> *Someone has not been looking at their subjects and they answered the Geography questions. […] But the person who answer the Geography question answered it correctly, so it deserved a well done for it. And then the other two [responses: "Since somebody answered a Geography question, I'm going to switch subjects." And "I should answer the Geography questions. Let's work on the subjects we chose."] were just going to create arguments,*

> *so we would waste time again. Probably I would have merged 'Nice work' with 'Stop answering my questions'.* (Anna)

> *I noticed that my subject is Geography, but I didn't answer that question […] I feel that the second one [response: "Whoever answered a Geography question, nice work!"] was more appropriate because it's encouraging them by complimenting them but in a way that they know they are doing well, so it's not giving them false hope, it's just saying that they are doing well, even though it's not their area to answer it. I think I would have preferred one that was more like, even though they answered an area that wasn't their own, it's still good. Even though it wasn't your subject area to answer it, you still got it right, so well done.* (Emily)

The alternative responses offered by the students are important to be considered here as well. If they had the option, students would both praise the teammate for getting a question correct and at the same time remind them of the agreed plan. Although students wanted to repair the shared understanding, the message supposed to do that was going to create arguments and disturb the team balance, and therefore it was not selected.

7.5.3 Communicating in real life

The consideration that participants seemed to give in real life to determine their responses indicates that there were more things to consider when responding in real life, than in the computer-based scenario they were involved. The selection of communication messages was determined by whether students would have said something similar in real life or not.

> *I don't think I would have said any of them [responses] if I was to say in real life.* (Pablo)

> *The answers were a struggle because they were four and because it's not your own words. You really have to think what each one means, because when it's sent through a chat to someone, the way it is send, the tone of*

> *your voice, it could mean completely different. So, you really have to think what each one actually means and whether you are being rational or you are being mean, that's pretty much what you have to think about, what you are actually saying.* (Stephan)

Students expressed their preference for wording their responses differently, making them less formal.

> *I would probably word them [responses] differently sometimes, just to keep them more relaxed. Instead of saying "the clock is ticking" just say "keep working you haven't got much time left".* (Pablo)

> *They [responses] were quite short and straight to the point. Probably I would word them quite differently and I would make my response longer. I would have been able to relate to them more.* (John)

> *The way I say things would be different because this was like formally typed, whereas I speak quite informally.* (Leo)

In addition, the assessment scenario does not allow for a humorous response or a joke, but as it is evident from students' interpretations there were instances in which they wanted to say something in humorous way or have interpreted something as a joke. For example, one student felt uncomfortable with the response options in item X6 and wanted to respond in a humorous way instead. When both teammates claimed the subject People, the student wanted to make a joke by claiming the People subject for herself.

> *It wasn't like super comfortable, it just felt a bit, like I am trying to lead them. […] I would say 'Take the other ones and I'll take People', but I probably wouldn't actually say that in a serious way, I would probably be joking.* (Anna)

Another student interpreted Zach's message in item X8 "*I guess Economy would be all right. I like money"* as a joke.

> *That one [Zach's message: "I guess Economy would be all right. I like money."] threw me off a bit, because I thought he was joking at first. When he said that, he doesn't understand economy it's not just money, he doesn't understand the financial side and all of it. So, looking back at it I think I should have said probably the response "Liking money doesn't mean you understand economy".* (Stephan)

Depending on how it was said in a real teenage-group problem-solving context, Zach's message might be interpreted as humorous. If it was a joke, the right response might be to support the humour in the situation. Also, some of the messages might have been interpreted by the students as a joke. For instance, giving the response "Liking money doesn't mean you understand economy" could have been interpreted by someone as a humorous thing to say in response to "I guess Economy would be all right. I like money". There is a sense in which the computer simulation removes everything except an assumption that students are computerised collaborative problem solvers. As a result of that, there is no humour or jokes, and students are just getting on with the test, which could be questioned whether it reflects real CPS.

7.5.4 Providing alternative responses

Students talked about their need to give an alternative response to the ones provided in the list of messages. It is obvious that there is a problem with this "technology" of assessment; it does not allow for the possibility of a student trying to open up a dialogue about an issue, but it tends to ask for the student to solve the problem in a "one-liner" instead. The question is, though, what if the best answer is really a pertinent question? For example, when Zach claimed the Economy subject in item X8 by saying "I guess Economy would be all right. I like money", students wanted to pose some questions to Zach, instead of claiming the remaining subject of Geography, which is the credited response for this item.

> *I could have said 'Why do you want to do Economy?'* (Becky)

> *I would probably have said 'Personally what are you more comfortable with, Geography or Economy? Don't feel rushed, but just make a decision on what you feel good, I'm happy to take either of that.'* (Stephan)

The fact that the student volunteers a response that they would like to have agreed with is important because they are not offered that option. There is no open response, and so when they are given the opportunity to give an open response, students actually give a good answer. Similarly, in item X11 students realised that Zach has not answered any questions in his assigned subject, and they wanted to ask him some questions. However, no question was offered as a response option in the list of messages.

> *I saw that he got no questions right, so I wanted to say, 'Are you struggling or something?'* (Oliver)

> *I would say, like, 'Do you need any help? Do you want to swap or something?'* (John)

> *I would say, 'How do you think it's going Zach? Which one would you find easier, Geography or Economy? Because we can always just switch.' Because I would be happy to just switch like that, if it would be easier for Zach.* (Stephan)

Another consequence of the limitations that students had in the chat space was that they felt uncomfortable for having to select one option that did not really reflect what they wanted to say. The student's response here suggests the need for another option such as "none of these", although this might be difficult to score later.

> *I didn't like any of the answers, it didn't feel as good putting an answer. I didn't feel I really want to put one to be fair.* (Stephan)

Having the option to type their response was also another point that was discussed as preferrable for students, so that they could provide a response that reflects what they wanted to say.

> *It's quite hard to give an answer if what I want to say in my head isn't an option, so I feel like typing would have been better because sometimes the answers weren't what I wanted to say.* (Emily)

7.5.5 Showing test savviness

The messages available to students for communication in the chat space were pre-determined, and all students were presented with the same four response options no matter what they have responded in every item. Students were found to be able to exclude options that were irrelevant with the situation or did not make sense. Being able to guess what the expected 'correct' response for the game is, raises issues about the validity of student responses.

> *Because there is no point talking about any of the other ones [responses]. The other ones are kind of pointless. It [response] is going to get the task done as quick, which is the challenge of the task.* (Anna)

> *They both want to take People and it's going to cause an argument over this one thing, so instead you need to think reasonably and think why do each one of them actually want this. So, I thought the rational response there. […] I thought well at the end of the day if I jumped and said the top one [response], that I want that subject, it would create an argument and divide the group. Then the third one [response], it doesn't resolve the issue, and the last one [response], I thought that wasn't helping either.* (Stephan)

> *I just thought what's the easiest way to answer it together. What would be the most efficient way completing the task? I feel like out of these options that was probably like the better way of figuring out how we are going to work as a team.* (Emily)

It was found that students could understand that they are in a test situation, and they understand what the examiners are getting at, e.g., saving time, avoiding arguments, resolving conflicts. For that reason, they get on with the test, showing test savviness, i.e., knowing what the examiner wants and giving it to them. In addition, as reported by the students, some of the response options in the list of pre-defined messages were obviously inappropriate, and therefore, it was easier to exclude them from the list.

> *The bottom three [responses: "Zach, aren't you the one who said we all had to work fast?", "Do you expect us to stop what we're doing and help you instead?", and "Are you behind because you were working on my Geography questions?"], they all come across very rude. There is no point in being rude when you can help someone.* (Emily)

> *I thought that the second response ["Zach, aren't you the one who said we all had to work fast?"] was a mean thing to say, because although he has been a bit hypocritical, it's not his fault he is struggling at the end of the day, and people struggle, the reason that he is struggling is because he is on the wrong subject.* (Stephan)

> *Saying that [response: "Since somebody answered a Geography question, I'm going to switch subjects."], it just seems like a childish response.* (Leo)

> *The last one [response: "Alice and Zach, are you going to answer questions faster than you choose subjects?"] sounded a bit immature to say.* (John)

7.5.6 Closing remarks on the results

The results presented in this section show that overall students suggested their responses were not authentic in the sense of how they would respond in the 'real' situation being simulated. Instead, they were mediated by the simulation, and more specifically, by a rationality that they understood the system to demand. This contradicted the affective and affiliative tone that they would prefer if the computer-simulated partners were real, embodied co-participants in the problem-solving situation.

7.6 Discussion

It has been recently argued that the extent to which the CPS skills assessed in the PISA 2015 CPS assessment represent the way students would interact with human partners, given the a priori constraints of the computer-simulated partner approach, needs to be determined (Herborn et al., 2020). Furthermore, it was argued that the complex nature of constructs such as CPS competence requires the investigation of vital quality evidence that goes beyond traditional analyses and is based on students' thinking and response processes (Ercikan & Oliveri, 2016). The current paper contributes to the investigation of those issues by analysing cognitive interview data from students interacting with the released PISA 2015 CPS task in some new ways that add to what PISA/OECD have reported to date.

Among the most important findings is the fact that students' responses were mediated by the simulation, and more specifically, by a rationality that they understood the system to demand. Therefore, their responses were not authentic in the sense of how they would respond in the 'real' situation being simulated. These findings confirm the points raised in Shaw and Child's (2017) critique of the PISA 2015 CPS assessment including the question of authenticity and whether there is a potential mismatch between how a student would respond in a natural setting and how they respond in the assessment.

It has been argued that it is unclear whether the pre-defined responses available to the student in the PISA 2015 CPS assessment were optimal, both relative to other responses, and to the infinite potential responses in a natural setting (Shaw & Child, 2017). Considering the evidence from the cognitive interview data, another important finding that sheds light to the point raised by Shaw and Child (2017) is that students could not relate to the responses offered in the lists of pre-defined messages. When they were given the opportunity to offer their own response, they wanted to ask questions and open dialogue with their teammates. Also, the formal tone and straight-to-the point messages did not allow them to use humour, for example, to repair understanding in the group. Students were also limited by the assessment technology in the chat space to selecting messages only after the computer-simulated partner was programmed to send a message first. This means that they could not initiate a conversation or send a message to the chat, unless it was in response to a message that they have received from the computer-simulated partners.

Computer-simulated scenario-based tasks have been previously argued to offer increased test-taker engagement, authenticity, and standardisation when compared to other assessment types, such as self-assessments, third-party evaluations, and observational tools, however the possibility of test-takers "gaming" the task persists (Oliveri et al., 2017, p. 21). An important finding of the current paper that contributes to above argument concerns demonstrating test savviness, or in other words, students showing awareness of what the task required them to do to progress and win the competition. In most of the items, students were able to guess which response was the 'correct' one by excluding response options that were obviously inappropriate or irrelevant.

Validity evidence presented in the current paper allows to formulate recommendations for future work on the development and use of computer-simulated CPS assessment tasks. It is suggested that this type of tasks should be used in combination with other types of assessment such as teacher evaluations and with the purpose of anchoring teacher's judgement rather than being the main instrument used for determining student performance. Test developers should focus on allowing for more freedom in the communication, one way to do that could be by including recorded video messages and, of course, real students as group members. One development towards that direction is the laboratory classroom Science of Learning Research Classroom at the University of Melbourne, which uses advanced video technology to capture simultaneous and continuous classroom social interactions using multiple cameras and microphones (Chan et al., 2018). The Social Unit of Learning project has used this laboratory classroom facility to examine individual, dyadic, small group and whole class problem solving and learning in mathematics (Chan & Clarke, 2017; Nieminen et al., 2022; Zhang et al., 2022). It has been argued that this facility made possible research designs that combine better approximation to natural social settings as well as conclusions about connections between interactive patterns of social negotiation and problem solving (Chan & Clarke, 2017).

Finally, researchers should collect more qualitative data in the form of students' retrospective reflections about their thought processes when responding to CPS tasks and use them as interpretative information and validity evidence. A clear message to policy makers is that the interaction with computer-simulated partners in scenario-based CPS tasks similar to the ones investigated in the present study render non-authentic student responses. Therefore, any efforts to incorporate CPS tasks as part of the curriculum or the national curriculum testing should be primarily focused on establishing good diagnostic instruments in regard to sources of invalidity such as adverse consequences.

7.6.1 Limitations

This paper also has some limitations that need to be addressed. Due to item confidentiality, only one PISA 2015 CPS task was available for public view and was therefore employed here to gather validity evidence. Another limitation is related to the sample size used, which was quite small (n = 10) and can be described as 'convenient sample'. However, the aim of the paper was not the representation of the population in the sample. In addition to that, CI as a method often requires small numbers of participants to allow more in-depth analysis. A third limitation concerns the cognitive interviewing process, which has been argued to be limited by the fact that not all cognitive processes can be verbalised by students (Collins, 2003). Finally, not all students might try to address the verbal probes they are asked during the CI based on what they really think, which can hinder the quality of the data collected. For that reason, the interview excerpts selected for presentation in the results include quotes in which students really explained their thinking process and reasoning.

7.7 Conclusion

This paper adds value to the existing literature and contributes to knowledge about the validity of PISA's CPS assessment in three ways. First, it offers an evaluation of evidence relevant to substantive aspect of validity by investigating the students' response processes. Second, it offers an evaluation of the external and consequential aspects of validity of the constructed measures by examining the authenticity of the assessment focussing on the pre-defined message options that students are presented with as a means to communicate with the computer-simulated partners. Finally, by highlighting aspects of validity that have not been targeted in the literature to date, this paper makes a significant contribution in informing policy and practice about the uses, limitations, and interpretations of the PISA 2015 CPS competence measure.

The main objective of this paper was to examine student response processes and offer critiques of the external validity of, primarily, the CPS competence assessment task, and in turn, the CPS framework operationalising CPS competence. Evidence for five categories of issues related to validity, namely identifying contextual concerns, showing emotional intelligence, communicating in real life, providing alternative responses, and showing test savviness, were found when comparing student response processes across items. Overall, CI provided useful validity evidence for the interpretation of the measurement derived from the PISA 2015 CPS task. So far, it has been debatable whether the level of control offered using the assessment technology adopted by PISA 2015 CPS task outweighs issues of and concerns about external and consequential validity. This paper presented evidence supporting the proposition that this sort of technology utilised in group problem-solving situations introduces limiting constraints to social interactions. Adopting such assessment technology might be considered advantageous for large-scale international assessments such as the PISA study, however, there are important sacrifices that need to be made concerning validity aspects. Whether such critical compromise is worthy is a matter of judgement that teachers, policy makers and researchers need to make considering the social consequences of the derived measurements.

7.8 References

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

Andrews-Todd, J., & Forsyth, C. M. (2020). Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Computers in Human Behavior*, *104*, 105759. https://doi.org/10.1016/j.chb.2018.10.025

Beatty, P. C., & Willis, G. B. (2007). Research Synthesis: The Practice of Cognitive Interviewing. *Public Opinion Quarterly*, *71*(2), 287–311. https://doi.org/10.1093/poq/nfm006

Benítez, I., & Padilla, J.-L. (2014). Analysis of Nonequivalent Assessments Across Different Linguistic Groups Using a Mixed Methods Approach: Understanding the Causes of Differential Item Functioning by Cognitive Interviewing. *Journal of Mixed Methods Research*, *8*(1), 52–68. https://doi.org/10.1177/1558689813488245

Castillo-Díaz, M., & Padilla, J.-L. (2013). How Cognitive Interviewing can Provide Validity Evidence of the Response Processes to Scale Items. *Social Indicators Research*, *114*(3), 963–975. https://doi.org/10.1007/s11205-012-0184-8

Chan, M. C. E., & Clarke, D. (2017). Structured affordances in the use of open-ended tasks to facilitate collaborative problem solving. *ZDM - Mathematics Education*, *49*(6), 951–963. https://doi.org/10.1007/s11858-017-0876-2

Chan, M. C. E., Clarke, D., & Cao, Y. (2018). The social essentials of learning: An experimental investigation of collaborative problem solving and knowledge construction in mathematics classrooms in Australia and China. *Mathematics Education Research Journal*, *30*(1), 39–50. https://doi.org/10.1007/s13394-017-0209-3

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research*, *12*(3), 229–238. https://doi.org/10.1023/A:1023254226592

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, *13*(1), 3–21. https://doi.org/10.1007/BF00988593

Creswell, J. W. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Sage.

De Boeck, P., & Scalise, K. (2019). Collaborative Problem Solving: Processing Actions, Time, and Performance. *Frontiers in Psychology*, *10*, 1280. https://doi.org/10.3389/fpsyg.2019.01280

Ercikan, K., & Oliveri, M. E. (2016). In Search of Validity Evidence in Support of the Interpretation and Use of Assessments of Complex Constructs: Discussion of Research on Assessing 21st Century Skills. *Applied Measurement in Education*, *29*(4), 310–318. https://doi.org/10.1080/08957347.2016.1209210

Glaser, B. G. (1978). *Theoretical Sensitivity: Advances in the Methocology of Grounded Theory* (first edition). Sociology Pr.

Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the Science of Collaborative Problem Solving. *Psychological Science in the Public Interest*, *19*(2), 59–92. https://doi.org/10.1177/1529100618808244

Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? *Computers in Human Behavior*, *104*, 105624. https://doi.org/10.1016/j.chb.2018.07.035

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A Framework for Teachable Collaborative Problem Solving Skills. In P. Griffin & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 37–56). Springer Netherlands. https://doi.org/10.1007/978-94-017-9395-7_2

Hopfenbeck, T. N., Lenkeit, J., Masri, Y. E., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research*, *62*(3), 333–353. https://doi.org/10.1080/00313831.2016.1258726

Hopfenbeck, T. N., & Maul, A. (2011). Examining Evidence for the Validity of PISA Learning Strategy Scales Based on Student Response Processes. *International Journal of Testing*, *11*(2), 95–121. https://doi.org/10.1080/15305058.2010.529977

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., Groot, E. D., Gilbert, M. C., Musu, L., Kempler, T. M., & Kelly, K. L. (2007). Cognitive Processing of Self-Report Items in Educational Research: Do They Think What We Mean? *Educational Psychologist*, *42*(3), 139–151. https://doi.org/10.1080/00461520701416231

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.

Messick, S. (1995). Validity of Psychological Assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.

Nieminen, J. H., Chan, M. C. E., & Clarke, D. (2022). What affordances do open-ended real-life tasks offer for sharing student agency in collaborative problem-solving? *Educational Studies in Mathematics*, *109*(1), 115–136. https://doi.org/10.1007/s10649-021-10074-9

Nouri, J., Åkerfeldt, A., Fors, U., & Selander, S. (2017). Assessing Collaborative Problem Solving Skills in Technology-Enhanced Learning Environments – The PISA Framework and Modes of Communication. *International Journal of Emerging Technologies in Learning (IJET)*, *12*(04), 163. https://doi.org/10.3991/ijet.v12i04.6737

OECD. (n.d.). *Description of the Released Unit from the 2015 PISA Collaborative Problem-Solving Assessment, Collaborative Problem-Solving Skills, and Proficiency Levels*. Retrieved 29 January 2022, from https://www.oecd.org/pisa/test/CPS-Xandar-scoring-guide.pdf

OECD. (2017a). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. OECD Publishing. https://doi.org/10.1787/9789264281820-en

OECD. (2017b). *PISA 2015 Results (Volume V): Collaborative problem solving*. OECD Publishing. https://doi.org/10.1787/9789264285521-en

OECD. (2017c). *PISA 2015 Technical Report*. OECD Publishing. https://doi.org/10.1787/9789264273856-19-en

Oliveri, M. E., Lawless, R., & Molloy, H. (2017). A Literature Review on Collaborative Problem Solving for College and Workforce Readiness. *ETS Research Report Series*, *2017*(1), 1–27. https://doi.org/10.1002/ets2.12133

Pepper, D., Hodgen, J., Lamesoo, K., Kõiv, P., & Tolboom, J. (2018). Think aloud: Using cognitive interviewing to validate the PISA assessment of student self-efficacy in mathematics. *International Journal of Research & Method in Education*, *41*(1), 3–16. https://doi.org/10.1080/1743727X.2016.1238891

Peterson, C. H., Peterson, N. A., & Powell, K. G. (2017). Cognitive Interviewing for Item Development: Validity Evidence Based on Content and Response Processes. *Measurement and Evaluation in Counseling and Development*, *50*(4), 217–223. https://doi.org/10.1080/07481756.2017.1339564

QSR International Pty Ltd. (2018). *NVivo* (Version 12) [Computer software]. https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

Scoular, C., Care, E., & Hesse, F. W. (2017). Designs for Operationalizing Collaborative Problem Solving for Automated Assessment. *Journal of Educational Measurement*, *54*(1), 12–35. https://doi.org/10.1111/jedm.12130

Scoular, C., Eleftheriadou, S., Ramalingam, D., & Cloney, D. (2020). Comparative analysis of student performance in collaborative problem solving: What does it tell us? *Australian Journal of Education*. https://doi.org/10.1177/0004944120957390

Shaw, S., & Child, S. (2017). Utilising technology in the assessment of collaboration: A critique of PISA's collaborative problem-solving tasks. *Research Matters: A Cambridge Assessment Publication*, *24*, 17–22.

Siddiq, F., & Scherer, R. (2017). Revealing the processes of students' interaction with a novel collaborative problem solving task: An in-depth analysis of think-aloud protocols. *Computers in Human Behavior*, *76*, 509–525. https://doi.org/10.1016/j.chb.2017.08.007

Strauss, A., & Corbin, J. M. (1998). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications.

Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education*, *143*, 103672. https://doi.org/10.1016/j.compedu.2019.103672

Tourangeau, R. (1984). Cognitive science and survey methods: Acognitive perspective. In T. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey design: Building a bridge between disciplines* (pp. 73–100). National Academy Press.

Willis, G. B. (2005). *Cognitive interviewing a tool for improving questionnaire design*. SAGE.

Zhang, S., Cao, Y., Chan, M. C. E., & Wan, M. E. V. (2022). A comparison of meaning negotiation during collaborative problem solving in mathematics between students in China and Australia. *ZDM – Mathematics Education*, *54*(2), 287–302. https://doi.org/10.1007/s11858-022-01335-9

Chapter 8 Discussion: contributions, limitations, and implications of this thesis

8.1 Significance and contribution to knowledge

This section presents the significance of the thesis by reflecting on what the state-of-the-art was in the field before and what it is now. This thesis consisted of four self-contained research papers presented in Chapters 4-7 with clear contributions to knowledge that are briefly presented here. However, it is more than a collection of the four research papers, and this section also presents what they all add up to in terms of the significant contribution to the field.

8.1.1 What was the state of the field before?

The recent educational literature proposed various collaborative problem solving (CPS) frameworks to inform the measurement of the CPS competence, however they were not systematically evaluated. The PISA 2015 CPS framework, for example, driven by its main interest to construct summative assessments to inform education systems, appears not to consider the concept of CPS from the different perspectives of individuals, groups, and communities (Cukurova, Luckin, Millán, et al., 2018; Fiore et al., 2010). Although valuable, these attempts still present some limitations. Several authors have pointed to the need for better definitions of CPS (e.g., Andrews-Todd & Kerr, 2019; Stadler, Shubeck, et al., 2020; Sun et al., 2020) as well as the need for consistency that will allow connection of new research with previous studies on both collaboration and problem solving.

Apart from defining a construct of interest, the various CPS frameworks have also aimed to inform test development of CPS assessment tasks. During the development of assessment instruments, researchers need to specify what areas of performance should be captured (Salas et al., 2017). To ensure that student performance is accurately assessed, the conceptual elements of a targeted construct should be clearly defined (Salas et al., 2017). A relevant question arising from reviewing the CPS frameworks is what criteria determined the skills that are measured. For example, in the framework for teachable CPS skills (Hesse et al., 2015) the detailed skills needed to be measurable in large-scale assessment, allow the derivation of behavioural indicators that can be assessed by teachers in classroom settings, and be teachable.

Other skills relevant to the CPS competence construct might have been excluded simply because they did not meet the above criteria. Additionally, practitioners might consider real, small-scale, authentic problem solving more beneficial for developing collaborations. Historically, the assessment of cognitive and social competencies has been claimed to rely on different approaches, i.e., correct versus incorrect answers for the assessment of cognitive competences and self-report estimates for the assessment of social competences (Care, Scoular, et al., 2016). More recently, efforts have been put forth to evaluate CPS skills themselves and design computer environments to support their measurement (Andrews-Todd & Forsyth, 2020).

Although there has been a recent increase in studies assessing students' CPS competence with the use of computer-based tasks, a systematic overview of the available literature has been lacking to date. Most of the previous educational assessment research about the validity of students' CPS competence measures, which derived from computer-based assessments such as the PISA 2015, is based on internal validity investigations, whereas aspects of validity such as structural, external, and consequential are less targeted. Currently, according to my systematic literature search there are no studies systematically assessing validity evidence based on student response processes in PISA's 2015 CPS assessment. In addition, since the release of PISA 2015 CPS results, hardly any studies have made use of the PISA 2015 CPS data or provided validity evidence for the test score interpretation and use (i.e., De Boeck & Scalise, 2019; Scoular, Eleftheriadou, et al., 2020; Tang et al., 2021). To the best of my knowledge there has been no exploration of the PISA 2015 CPS data for England in which structural and external validity aspects are examined.

8.1.2 Contribution to knowledge made by Chapter 4

Chapter 4 (CPS concepts review) contributes to our understanding of CPS in two ways: (i) by focusing systematically on the conceptualisation of CPS within recent educational research evidence, and (ii) by looking at a range of educational contexts from primary to higher education, extending previous reviews.

A total of 59 articles were deemed relevant for review and the analysis led to three categories of conceptualisations: CPS competence, CPS practice, and CPS interaction. Articles adopting a CPS competence focus (n = 36) emphasised individual capacities within collaborative situations, as opposed to focusing on the collaborating pair/group. They focused on assessing CPS as an intended learning outcome, making assumptions about the individual (student's) CPS competence underlying their observed behaviours. In CPS practice (n = 17), authors generally emphasised the individual cognition, investigating how it was affected by cultural mediations, or even collaborative interactions. CPS was then conceived as a pedagogical approach, providing the social context in which learning could take place. Finally, in CPS interaction (n = 6), CPS was conceived as taking place in a negotiated and shared conceptual space in which individual student contributions could not be distinguished from the group co-constructed meaning making process.

I argue that the study of how CPS is both conceptualised and operationalised can provide valuable information for critically situating the concept which researchers opt to use. The review identified the strengths and weaknesses of CPS-competence, CPS-practice, and CPS-interaction concepts and models and argued that future research should ensure to situate their work in regards to these existing categories to maintain conceptual coherence. It is suggested that the coherent use of such conceptualisations will help overcome previous problems in CPS research lacking in coherence and consistency and help research become more cumulative for the research field itself, but also for policy and practice.

Chapter 4 also informs policy makers about the dangers of backwash of assessment on pedagogy and the need for research to move through other methodologies that complement the competence-focused view of CPS. It therefore argues for a focus on how the collective can make meaningful progress in problem solving and learning. In addition, this paper suggests that more information about the validity and authenticity of the individualised assessments of CPS competence is needed and advocates future research studies to focus on the exploration of CPS from different levels of description, extending existing conceptual frames and in relation to the purpose of research and its epistemology.

The findings of Chapter 4 will also be of interest to scholars researching CPS since the comprehensive overview presented can serve as a starting point for future reviews and for those who are new to CPS research.

8.1.3 Contribution to knowledge made by Chapter 5

Chapter 5 (CPS measurement review) adds value to the existing literature and contributes to knowledge about the assessment of CPS competence in three ways. First, it offers a systematic selection and categorisation of the state-of-the-art of CPS competence assessment, especially in relation to the computer-simulation and educational measurement subfield. It therefore makes a significant contribution in informing future research (including future reviews and meta-analyses). This review includes 26 articles reporting 15 assessments, most of which target secondary students working in groups of two and communicating through text messages.

Second, this paper offers an evaluation of the relationship between conceptualisation of CPS competence and assessment instruments. By highlighting current facets of CPS competence that are inadequately covered, namely active listening, audience awareness, team empowerment, and team learning, it informs the development of more comprehensive assessment instruments. Third, this paper offers an evaluation of the strategies used to validate existing CPS competence measures. By highlighting the types of validity evidence that have been overlooked (i.e., external, substantive, and consequential validity), the development of future validation studies is facilitated. Chapter 5 suggests the potential of researching students in real-life situations to inform assessment development. In this way, the derived measures will subsequently be more authentically well aligned with CPS in authentic situations.

It is suggested that future research studies should focus on analytical methods that can exploit the rich information captured in student communication. To move the research on the assessment of CPS competence forward, this paper advocates for small-scale research designs that investigate students in authentic situations. In this way, evidence for the validity aspects that is currently lacking (i.e., external) could be evaluated.

Finally, the findings of Chapter 5 will be of interest to researchers involved in the development of computer-simulated (CPS) assessments and validation, since they summarise what has been already measured and how, pointing to current limitations and ways for improvement.

8.1.4 Contribution to knowledge made by Chapter 6

As suggested previously, the validity of CPS competence measures developed using computer-simulated, scenario-based assessments, such as the PISA 2015 CPS assessment, has been an area of interest in current literature. Nevertheless, my systematic literature review (Chapter 5) found no articles using PISA 2015 CPS data for England to explore validity issues.

Chapter 6 (Rasch analysis of PISA 2015 CPS measure and its correlates with important variables) contributes to knowledge about the validity of PISA's CPS competence measure in three ways. First, it offers an evaluation of evidence relevant to structural aspect of validity by investigating the measurement properties and the hypothesised multi-dimensional structure of the measure. The hypothesised dimensions are the three newly conceptualised competencies of CPS (i.e., 'establishing and maintaining shared understanding', 'taking appropriate action to solve the problem', and 'establishing and maintaining team organisation') as found in PISA's theoretical framework for CPS (OECD, 2017a). For this analytical sample and set of items, the three sub-scales showed similar results in their associations with other variables analysed, and as compared to the overall CPS competence measure. It was therefore concluded that CPS items collectively measure a unidimensional model, presumed to be of CPS competence.

Second, Chapter 6 offers an evaluation of the external and consequential aspects of validity of the constructed measures by examining their association with theoretically relevant constructs. The CPS competence measures were only very weakly correlated with students' attitudes towards collaboration. The fact that CPS competence measures (overall and sub-scales) were found to be highly correlated with performance in other subject domains might suggest that the cognitive aspect is more prevalent than the collaborative aspect in the social constructs developed.

On the other hand, it might be argued that individual students' competence in, for instance, science is the result of successful learning through collaboration with others, and to this extent a good indicator of CPS, or at least as good as their self-reported attitudes to collaboration in a questionnaire context.

By highlighting aspects of validity that have not been targeted in the literature to date (i.e., external, and consequential validity), Chapter 6 makes a significant contribution by informing policy and practice about the uses, limitations, and interpretations of the PISA CPS competence measure. One of the most important messages that the paper conveys is that there is still a question about whether measures derived from such controlled tests have external validity regarding students' capabilities to work together with others effectively on problem solving in reality.

Publishing of the PISA 2015 CPS results will likely increase the attention received from researchers, educators, and policy makers on students' CPS competence and this, in turn, might drive policy decisions and curricula revisions. Therefore, the findings of this chapter will also be of interest to policy makers who are encouraged to be more sensitive to external and consequential validity considerations as well as to scholars involved in validation work.

8.1.5 Contribution to knowledge made by Chapter 7

Chapter 7 (Cognitive interviewing of students) sheds light into how students understand the PISA 2015 CPS items and how they respond to them. I argue that this in-depth analysis of student response processes makes it possible to identify points of conflict between the PISA 2015 CPS framework and students' motivation for selecting certain responses. Given the continued influence of the PISA study and the general lack of validity evidence for CPS competence assessments, Chapter 7 suggests the potential for cognitive interviewing as a method for assessing validity evidence based on content and response processes. This method, although highly relevant for the validation of CPS competence measures, has been surprisingly neglected in the literature so far.

To the best of my knowledge there has been no study, up to the writing of this thesis, exploring the validity of the PISA 2015 CPS competence measure through cognitive interviews. Chapter 7 offers a more comprehensive validation of the PISA 2015 CPS assessment. Results of this chapter add value to the existing literature and contribute to knowledge about the validity of PISA's CPS competence measure in three ways. First, it offers an evaluation of evidence relevant to the substantive aspect of validity by investigating the students' response processes. Following grounded theory for data analysis, five categories of student response processes were found: 'identifying contextual concerns', 'showing emotional intelligence', 'communicating in real life', 'providing alternative responses', and 'showing test savviness'. Second, it offers an evaluation of the external and consequential aspects of validity of the constructed measures by examining the authenticity of the assessment focussing on the pre-defined message options that students are presented with as a means to communicate with the computer-simulated partners.

Overall, Chapter 7 shows that student responses were mediated by the simulation, and more specifically, by a rationality that they understood the system to demand, which contradicted the affective and affiliative tone that they would prefer if the computer-simulated partners were real, embodied co-participants in the problem-solving situation. Finally, by highlighting aspects of validity that have not been targeted in the literature to date (i.e., substantive validity), this chapter makes a significant contribution in informing policy and practice about the uses, limitations, and interpretations of the PISA 2015 CPS competence measure. As argued previously, it also advocates for the need of educational policy to be sensitive to the authenticity of assessments and the social consequences of test score use and interpretation.

The findings of Chapter 7 will also be of interest to survey methodologists and educational researchers and to scholars using cognitive interviewing to investigate validity evidence.

8.1.6 What do they all add up to?

The thesis provides a validation of the PISA 2015 CPS competence assessment using a unified definition of validity. Drawing on the results of the self-contained papers, it provides a constructive critique of the PISA 2015 CPS assessment and theoretical framework, with implications for future CPS competence assessments more generally. Establishing validity entails taking into consideration multiple sources of relevant information (Karabenick et al., 2007; Messick, 1995; Wolfe & Smith, 2007a). Employing a sequential mixed-methods approach, this thesis provides (qualitative and quantitative) empirical evidence (Chapter 6 and 7) informed by two systematic literature reviews (Chapter 4 and 5), which are primarily qualitative, being conceptual, before getting into developing quantitative summaries.

Overall, the thesis provides a well-documented approach to validation which combines (i) conceptual and methodological analyses based on literature reviews of CPS competence, (ii) secondary data analysis of the PISA 2015 dataset; with (iii) primary CI data and analyses, which researchers can draw on for further validations. By using a unified definition of validity to inform the validation of the measures, this thesis will make a significant contribution in the development of future validation studies.

Specifically, the thesis proposes an approach to validation in the context of mixed methods research, motivated by the inadequacy of the methodologies used so far. One novelty of this thesis is the application of a mixed methods research design that combines statistical methods and CI to study validity evidence. The research design combined an exploration of the multidimensional character of CPS competence (structural validity) and its association with supposedly relevant constructs (external and consequential validity) with interpretations made by the students (substantive, external, consequential validity).

Among the most important findings of the thesis is the identification of student response processes from the cognitive interviews that suggests limits to the validity of the items that undermine the external validity. Adopting assessment technology, such as the PISA 2015 CPS assessment, might be considered advantageous for large-scale international assessments such as the PISA study, however, there are important sacrifices that need to be made concerning validity aspects. Whether such critical compromise is worthy is a matter of judgement that teachers, policy makers and researchers need to make considering the social consequences of the derived measurements.

8.2 Limitations of thesis

Limitations related to the four research papers will not be repeated here as they were covered in detail in Chapters 4-7. In this section, I re-contextualise them in light of the validity evidence examined in the thesis. Specifically, the problems of defining CPS coherently (discussed in Chapters 4 and 5) make measuring it risky, and this influences all the concerns that the thesis revealed empirically with (construct) validity.

Addressing different research questions with varied methodological perspectives allowed answering multiple and diverse research questions in the context of one study. For example, a limitation of Chapter 6 was the lack of substantive validity evidence since student response processes were not provided in the PISA 2015 dataset. Substantive validity is one of the main validity aspects in the unified validity framework (Messick, 1995) and, therefore, it was important to be investigated to determine whether the theoretical processes are engaged by respondents in the assessment tasks. This limitation was addressed in Chapter 7, where student response processes were an essential part of the research design. These were investigated through cognitive interviews and more specifically drawing from students' retrospective reflections on explaining their reasons for arriving at a response.

There are some unavoidable limitations due to the methods used in this thesis. Specifically, student interview data are subject to satisficing, which occurs when the respondent simply provides an answer without trying to address the question. An additional limitation relates to the fact that no observational data were analysed to explore external validity of the CPS competence assessment. However, analysing data from lesson observations in which students solve problem in groups in their classroom, in addition to completing the PISA 2015 CPS tasks, could give more insight into the subject. Finally, the most critical limitation that shaped the focus of this study relates to item confidentiality. In particular, the availability of the CPS items' content (or the absence of it) has posed certain challenges in the independent use of the PISA 2015 data.

Only a small proportion (about 10%) of CPS assessment items was released by test constructors for public view (i.e., the Xandar items), meaning that users of the data must rely upon the descriptions of the test instrument provided. As pointed out by Baird et al. (2017), without sight of the items, the data can only be interpreted through the lenses of those who constructed the test. PISA 2015 dataset does not include information about the possible actions that students make. Therefore, it is not possible to examine what is happening during the task for students who have made an 'incorrect' action. CPS competence was an innovative domain for PISA 2015, and therefore, there is no expectation for the material to be used in the following cycles again. Hence, there is less call for secure items as compared to the recurring domains (i.e., reading, science and mathematics). As stated previously, this limitation was somewhat addressed in Chapter 7 with the collection of student response processes for the available CPS assessment items.

8.3 Further research

Future research studies need to carefully develop their conceptualisations considering existing conceptual frames and in relation to the purpose of research and its epistemology. It is argued that correspondence between the conceptualisation of CPS and the unit of analysis employed to operationalise the construct is considered necessary for the research community to move towards a conceptual coherence in CPS-related research.

So far, the literature has been limited to examining differences in student responses when communicating with computer-simulated partners versus real humans. It is therefore suggested that researchers should focus on more in-depth analysis of student responses in real problem-solving situations. In addition, future developments in task design may target facilitating more authentic CPS competence assessments. Test developers should focus on allowing for more freedom in the communication; one way to do that could be by including recorded video messages and, of course, real students as group members. Furthermore, future research needs to exploit the rich information that the content of communication contains as part of measurement. To this end, alternative methods of scoring such as comparative judgement, previously proved to be useful for scoring difficult-to-define constructs (e.g., Jones et al., 2015; Jones & Inglis, 2015), could be applied in the context of CPS.

To get more comprehensive and authentic measures of students' CPS competence, what is needed from future research is a more holistic understanding of why students respond in the way they do in the collaborative assessments and what aspects need to be scored. Given the complexity of CPS competence concept, prioritising some (possibly small-scale) qualitative studies, where students explain their response processes, could help get that in-depth analysis. In addition, future research should focus on evaluating evidence for the validity aspects that have not been fully covered in the literature yet. Finally, results from this study are relevant to the specific sample of students in England, and therefore, future research needs to check whether they are applicable in other countries and cultures. Comparability issues for PISA 2015 CPS data might arise if similar validity issues are found to apply in other countries.

8.4 Implications for policy and practice

Results from this thesis have implications for educational policy and general research practice.

Educational policy should be more sensitive to external and consequential validity considerations. Specifically, policy that typically focuses on results from large-scale international surveys to inform curriculum changes and educational reform regarding students' CPS competence should consider the consequences of using such standardised CPS assessments, following their limitations.

Policy makers might attempt to include CPS competence assessments in the curriculum, especially after the publication of PISA results. What needs to be carefully considered though, are the consequences of introducing PISA-like CPS assessment tasks as high-stakes assessment or as part of the curriculum in schools. There is a risk that teachers would teach students to the test, which means getting students to practise artificial exercises, so that they can score high marks. There is also a risk that, if teachers rely exclusively on constrained assessment environments, then students might build up unrealistic expectations of what authentic collaboration might be in practice. Policy makers need to be aware of those dangers to prevent assessment driving this (collaborative problem solving) curriculum rather than supporting it.

Another policy implication is the need to use teacher assessment and the performance of the whole group in combination to individual student assessments of CPS competence. The qualitative work has shown that there could be a role for teacher assessment and the observation of whole group performance in real life classroom situation, and so the individual assessment could be used as an anchor or moderating instrument to complement the teacher assessment. Using a combination of teacher assessment and individual student assessments might be a step forward.

Chapter 9 References

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

Adams, R., Weale, S., Bengtsson, H., & Carrell, S. (2016, December 6). UK schools fail to climb international league table. *The Guardian*. https://www.theguardian.com/education/2016/dec/06/english-schools-core-subject-test-results-international-oecd-pisa

Albert, L., & Kim, R. (2013). Developing Creativity Through Collaborative Problem Solving. *Journal of Mathematics Education at Teachers College*, *4*(Fall-Winter), 32–38.

Alexander, P. A. (2020). Methodological Guidance Paper: The Art and Science of Quality Systematic Reviews. *Review of Educational Research*, *90*(1), 6–23. https://doi.org/10.3102/0034654319854352

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

Ananiadou, K., & Claro, M. (2009). *21st Century Skills and Competences for New Millennium Learners in OECD Countries*. OECD. https://doi.org/10.1787/218525261154

Anders, J., Has, S., Jerrim, J., Shure, N., & Zieger, L. (2021). Is Canada really an education superpower? The impact of non-participation on results from PISA 2015. *Educational Assessment, Evaluation and Accountability*, *33*(1), 229–249. https://doi.org/10.1007/s11092-020-09329-5

Andrews-Todd, J., & Forsyth, C. M. (2020). Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Computers in Human Behavior*, *104*, 105759. https://doi.org/10.1016/j.chb.2018.10.025

Andrews-Todd, J., Jackson, G. T., & Kurzum, C. (2019). Collaborative Problem Solving Assessment in an Online Mathematics Task. *ETS Research Report Series*, *2019*(1), 1–7. https://doi.org/10.1002/ets2.12260

Andrews-Todd, J., & Kerr, D. (2019). Application of Ontologies for Assessing Collaborative Problem Solving Skills. *International Journal of Testing*, *19*(2), 172–187. https://doi.org/10.1080/15305058.2019.1573823

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The Skill Content of Recent Technological Change: An Empirical Exploration. *The Quarterly Journal of Economics*, *118*(4), 1279–1333. https://doi.org/10.1162/003355303322552801

Avvisati, F., & Keslair, F. (2014). *REPEST: Stata module to run estimations with weighted replicate samples and plausible values* (revised 05 Jun 2019)

[Computer software]. Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s457918.html

Bailey, C., Tully, M. P., Pampaka, M., & Cooke, J. (2017). Rasch analysis of the Antimicrobial Self-Assessment Toolkit for National Health Service (NHS) Trusts (ASAT v17). *The Journal of Antimicrobial Chemotherapy*, *72*(2), 604–613. https://doi.org/10.1093/jac/dkw434

Baird, J.-A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: Fields apart? *Assessment in Education: Principles, Policy & Practice*, *24*(3), 317–350. https://doi.org/10.1080/0969594X.2017.1319337

Baird, J.-A., Isaacs, T., Johnson, S., Stobart, G., Yu, G., Sprague, T., & Daugherty, R. (2011). *Policy effects of PISA*. Oxford University Centre for Educational Assessment. http://oucea.education.ox.ac.uk/wordpress/wp-content/uploads/2011/10/Policy-Effects-of-PISA-OUCEA.pdf

Baird, J.-A., Johnson, S., Hopfenbeck, T. N., Isaacs, T., Sprague, T., Stobart, G., & Yu, G. (2016). On the supranational spell of PISA in policy. *Educational Research*, *58*(2), 121–138. https://doi.org/10.1080/00131881.2016.1165410

Beatty, P. C. (2004). The Dynamics of Cognitive Interviewing. In *Methods for Testing and Evaluating Survey Questionnaires* (pp. 45–66). Wiley-Blackwell. https://doi.org/10.1002/0471654728.ch3

Beatty, P. C., & Willis, G. B. (2007). Research Synthesis: The Practice of Cognitive Interviewing. *Public Opinion Quarterly*, *71*(2), 287–311. https://doi.org/10.1093/poq/nfm006

Benítez, I., & Padilla, J.-L. (2014). Analysis of Nonequivalent Assessments Across Different Linguistic Groups Using a Mixed Methods Approach: Understanding the Causes of Differential Item Functioning by Cognitive Interviewing. *Journal of Mixed Methods Research*, *8*(1), 52–68. https://doi.org/10.1177/1558689813488245

Berezner, A., & Adams, R. J. (2017). Why Large-Scale Assessments Use Scaling and Item Response Theory. In P. Lietz, J. C. Cresswell, K. F. Rust, & R. J. Adams (Eds.), *Implementation of Large-Scale Education Assessments* (pp. 323–356). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118762462.ch13

Bieber, T., & Martens, K. (2011). The OECD PISA Study as a Soft Power in Education? Lessons from Switzerland and the US: European Journal of Education, Part I. *European Journal of Education*, *46*(1), 101–116. https://doi.org/10.1111/j.1465-3435.2010.01462.x

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). Springer Netherlands.

Birnbaum, A. (1968). Some latent trait models and their use in inferring a student’s ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Addison-Wesley.

Boeije, H., & Willis, G. (2013). The Cognitive Interviewing Reporting Framework (CIRF): Towards the harmonization of cognitive testing reports. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(3), 87–95. https://doi.org/10.1027/1614-2241/a000075

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Lawrence Erlbaum.

Bourdieu, P. (1977). *Outline of a theory of practice*. University Press.

Breakspear, S. (2012). *The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance*. OECD. https://doi.org/10.1787/5k9fdfqffr28-en

Brussow, J. A. (2018). Consequential Validity Evidence. In B. B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (pp. 372–374). SAGE Publications, Inc. https://doi.org/10.4135/9781506326139

Cáceres, M., Nussbaum, M., Marroquín, M., Gleisner, S., & Marquínez, J. T. (2018). Building arguments: Key to collaborative scaffolding. *Interactive Learning Environments*, *26*(3), 355–371. https://doi.org/10.1080/10494820.2017.1333010

Cai, H., Lin, L., & Gu, X. (2016). Using a semantic diagram to structure a collaborative problem solving process in the classroom. *Educational Technology Research and Development*, *64*(6), 1207–1225. https://doi.org/10.1007/s11423-016-9445-6

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

Camacho-Morles, J., Slemp, G. R., Oades, L. G., Morrish, L., & Scoular, C. (2019). The role of achievement emotions in the collaborative problem-solving performance of adolescents. *Learning and Individual Differences*, *70*, 169–181. https://doi.org/10.1016/j.lindif.2019.02.005

Camara, W., O'Connor, R., Mattern, K., & Hanson, M. A. (2015). *Beyond Academics: A Holistic Framework for Enhancing Education and Workplace Success* (Act Research Report Series). ACT Inc.

Care, E., Anderson, K., & Kim, H. (2016). *Visualizing the Breadth of Skills Movement Across Education Systems*. Center for Universal Education at the Brookings Institution.

Care, E., & Griffin, P. (2014). An Approach to Assessment of Collaborative Problem Solving. *Research and Practice in Technology Enhanced Learning*, *9*(3), 367–388.

Care, E., Griffin, P., & Wilson, M. (2018). *Assessment and Teaching of 21st Century Skills: Research and Applications* (1st edition..). Springer International Publishing. https://doi.org/10.1007/978-3-319-65368-6

Care, E., Scoular, C., & Griffin, P. (2016). Assessment of Collaborative Problem Solving in Education Environments. *Applied Measurement in Education*, *29*(4), 250–264. https://doi.org/10.1080/08957347.2016.1209204

Caro, D. H., Sandoval-Hernández, A., & Lüdtke, O. (2014). Cultural, social, and economic capital constructs in international assessments: An evaluation using exploratory structural equation modeling. *School Effectiveness and School Improvement*, *25*(3), 433–450. https://doi.org/10.1080/09243453.2013.812568

Cascella, C., & Pampaka, M. (2020). Attitudes towards gender roles in family: A Rasch-based validation study. *Journal of Applied Measurement*. https://www.research.manchester.ac.uk/portal/en/publications/attitudes-towards-gender-roles-in-family-a-raschbased-validation-study(d552879e-6a1f-4347-9c57-54e5602e5ed5).html

Castillo-Díaz, M., & Padilla, J.-L. (2013). How Cognitive Interviewing can Provide Validity Evidence of the Response Processes to Scale Items. *Social Indicators Research*, *114*(3), 963–975. https://doi.org/10.1007/s11205-012-0184-8

Chan, M. C. E., & Clarke, D. (2017). Structured affordances in the use of open-ended tasks to facilitate collaborative problem solving. *ZDM - Mathematics Education*, *49*(6), 951–963. https://doi.org/10.1007/s11858-017-0876-2

Chan, M. C. E., Clarke, D., & Cao, Y. (2018). The social essentials of learning: An experimental investigation of collaborative problem solving and knowledge construction in mathematics classrooms in Australia and China. *Mathematics Education Research Journal*, *30*(1), 39–50. https://doi.org/10.1007/s13394-017-0209-3

Chang, C.-J., Chang, M.-H., Chiu, B.-C., Liu, C.-C., Fan Chiang, S.-H., Wen, C.-T., Hwang, F.-K., Wu, Y.-T., Chao, P.-Y., Lai, C.-H., Wu, S.-W., Chang, C.-K., & Chen, W. (2017). An analysis of student collaborative problem solving activities mediated by collaborative simulations. *Computers and Education*, *114*(C), 222–235. https://doi.org/10.1016/j.compedu.2017.07.008

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.

Cobb, P., & Gravemeijer, K. (2008). Experimenting to Support and Understand Learning Processes. In A. E. Kelly, R. A. Lesh, & J. Y. Baek (Eds.), *Handbook of Design Research Methods in Education*. Routledge.

Cohen, E. G. (1994). Restructuring the Classroom: Conditions for Productive Small Groups. *Review of Educational Research*, *64*(1), 1–35. https://doi.org/10.3102/00346543064001001

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research*, *12*(3), 229–238. https://doi.org/10.1023/A:1023254226592

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, *13*(1), 3–21. https://doi.org/10.1007/BF00988593

Coughlan, S. (2019, December 3). Pisa tests: UK rises in international school rankings. *BBC News*. https://www.bbc.com/news/education-50563833

Creese, B., Gonzalez, A., & Isaacs, T. (2016). Comparing international curriculum systems: The international instructional systems study. *The Curriculum Journal*, *27*(1), 5–23. https://doi.org/10.1080/09585176.2015.1128346

Creswell, J. W. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Sage.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches* (2nd ed.). Sage.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. SAGE Publications.

Cukurova, M., Luckin, R., & Baines, E. (2018). The significance of context for the emergence and implementation of research evidence: The case of collaborative problem-solving. *Oxford Review of Education*, *44*(3), 322–337. https://doi.org/10.1080/03054985.2017.1389713

Cukurova, M., Luckin, R., Millán, E., & Mavrikis, M. (2018). The NISPI framework: Analysing collaborative problem-solving from students' physical interactions. *Computers & Education*, *116*, 93–109. https://doi.org/10.1016/j.compedu.2017.08.007

De Boeck, P., & Scalise, K. (2019). Collaborative Problem Solving: Processing Actions, Time, and Performance. *Frontiers in Psychology*, *10*, 1280. https://doi.org/10.3389/fpsyg.2019.01280

Denzin, N. K., & Lincoln, Y. S. (Eds.). (1994). *Handbook of qualitative research* (pp. xii, 643). Sage Publications, Inc.

Department for Education. (2017). *Achievement of 15-Year-Olds in England: PISA 2015 Collaborative Problem Solving National Report* [Research brief]. Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/661243/PISA_2015_CPS_National_Report_FINAL.pdf

Dewey, J. (1910). *How we think* (pp. vi, 228). D C Heath. https://doi.org/10.1037/10903-000

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process* (revised edition). D.C. Heath and company.

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

Dillenbourg, P., Baker, M. J., Blaye, A., & O'Malley, C. (1996). The evolution of research on collaborative learning. In E. Spada & P. Reiman (Eds.), *Learning in Humans and Machine: Towards an interdisciplinary learning science* (pp. 189–211). Elsevier.

Dillenbourg, P., & Traum, D. (2006). Sharing Solutions: Persistence and Grounding in Multimodal Collaborative Problem Solving. *Journal of the Learning Sciences*, *15*(1), 121–151. https://doi.org/10.1207/s15327809jls1501_9

Ding, N. (2009). Visualizing the sequential process of knowledge elaboration in computer-supported collaborative problem solving. *Computers & Education*, *52*(2), 509–519. https://doi.org/10.1016/j.compedu.2008.10.009

Dominowski, R. L., & Bourne, L. E. (1994). History of Research on Thinking and Problem Solving. In R. J. Sternberg (Ed.), *Thinking and Problem Solving* (Vol. 2, pp. 1–35). Academic Press.

El Masri, Y. H., Baird, J.-A., & Graesser, A. (2016). Language effects in international testing: The case of PISA 2006 science items. *Assessment in Education: Principles, Policy & Practice*, *23*(4), 427–455. https://doi.org/10.1080/0969594X.2016.1218323

Eleftheriadou, S. (2019a). *An exploration of student responses to the PISA 2015 collaborative problem-solving assessment: A mixed-methods approach*. SEED PGR Conference 2019, The University of Manchester, 21 May 2019.

Eleftheriadou, S. (2019b). *An exploration of student responses to the PISA 2015 collaborative problem-solving assessment: A mixed-methods approach*. Methods X Conference 2019, The University of Liverpool, 17 May 2019.

Eleftheriadou, S. (2019c). *Conceptualisation and measurement of collaborative problem solving*. British Educational Research Association (BERA) Annual Conference 2019, The University of Manchester, 10-12 September 2019.

Eleftheriadou, S. (2019d). *Conceptualisation and measurement of collaborative problem solving: A systematic review of the literature*. Emerging Researchers' Conference 2019, Universität Hamburg, 2-3 September 2019.

Eleftheriadou, S. (2021). *Examining Construct Representation and Dimensionality of the PISA 2015 Collaborative Problem-Solving Measure*. European Conference for Educational Research (ECER) 2021, The University of Geneva (online), 6-10 September 2021.

Eleftheriadou, S., & Pampaka, M. (2019). *Examining evidence for the validity of PISA 2015 collaborative problem-solving measure using the Rasch model*. Australian Association for Research in Education (AARE) Conference 2019, Queensland University of Technology, 1-5 December 2019.

Engeström, Y. (1999). Activity Theory and Individual and Social Transformation. In Y. Engeström, R. Miettinen, & R.-L. Punamäki-Gitai (Eds.), *Perspectives on Activity Theory* (pp. 19–38). Cambridge University Press.

Ercikan, K., & Oliveri, M. E. (2016). In Search of Validity Evidence in Support of the Interpretation and Use of Assessments of Complex Constructs: Discussion of Research on Assessing 21st Century Skills. *Applied Measurement in Education*, *29*(4), 310–318. https://doi.org/10.1080/08957347.2016.1209210

Ercikan, K., Roth, W.-M., & Asil, M. (2015). Cautions about Inferences from International Assessments: The Case of PISA 2009. *Teachers College Record*, *117*(1).

Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data* (Revised Edition). A Bradford Book.

Farra, H. (1988). The Reflective Thought Process: John Dewey Re-visited. *The Journal of Creative Behavior*, *22*(1), 1–8. https://doi.org/10.1002/j.2162-6057.1988.tb01338.x

Fernandez-Cano, A. (2016). A methodological critique of the PISA evaluations. *RELIEVE*, *22*(1), 1–16.

Fiore, S. M., Graesser, A., Greiff, S., Griffin, P., Gong, B., Kyllonen, P., Massey, C., O'Neil, H., Pellegrino, J., Rothman, R., Soulé, H., & von Davier, A. (2017). *Collaborative Problem Solving: Considerations for the National Assessment of Educational Progress*. National Center for Education Statistics. http://orbilu.uni.lu/handle/10993/31897

Fiore, S. M., Rosen, M. A., Smith-Jentsch, K. A., Salas, E., Letsky, M., & Warner, N. (2010). Toward an understanding of macrocognition in teams: Predicting processes in complex collaborative contexts. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(2), 203–224. https://doi.org/10.1177/0018720810369807

Gao, Q., Zhang, S., Cai, Z., Liu, K., Hui, N., & Tong, M. (2022). Understanding student teachers' collaborative problem solving competency: Insights from process data and multidimensional item response theory. *Thinking Skills and Creativity*, *45*, 101097. https://doi.org/10.1016/j.tsc.2022.101097

Ghaderi, I. (2018). Unitary View of Validity. In B. B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (pp. 1752–1756). SAGE Publications, Inc. https://doi.org/10.4135/9781506326139

Gillies, R. (2016). Cooperative Learning: Review of Research and Practice. *Australian Journal of Teacher Education*, *41*(3). https://doi.org/10.14221/ajte.2016v41n3.3

Glaser, B. G. (1978). *Theoretical Sensitivity: Advances in the Methocology of Grounded Theory* (first edition). Sociology Pr.

Glaser, B. G. (1992). *Basics of Grounded Theory Analysis: Emergence Vs. Forcing* (1st edition). Sociology Pr.

Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 319–330. https://doi.org/10.1080/0969594042000304618

Goldstein, H. (2017). Measurement and evaluation issues with PISA. In L. Volante (Ed.), *The PISA effect on global educational governance* (pp. 49–58). Routledge.

Gough, D., Oliver, S., & Thomas, J. (2016). *An introduction to systematic reviews* (2nd ed.). SAGE Publications Ltd.

Graesser, A. C., Cai, Z., Morgan, B., & Wang, L. (2017). Assessment with computer agents that engage in conversational dialogues and trialogues with learners. *Computers in Human Behavior*, *76*, 607–616. https://doi.org/10.1016/j.chb.2017.03.041

Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the Science of Collaborative Problem Solving. *Psychological Science in the Public Interest*, *19*(2), 59–92. https://doi.org/10.1177/1529100618808244

Graesser, A. C., Greiff, S., Stadler, M., & Shubeck, K. T. (2020). Collaboration in the 21st century: The theory, assessment, and teaching of collaborative problem solving. *Computers in Human Behavior*, *104*, 106134. https://doi.org/10.1016/j.chb.2019.09.010

Grant, P. E., Pampaka, M., Payne, K., Clarke, A., & McAllister, M. (2019). Developing a short-form of the Genetic Counselling Outcome Scale: The Genomics Outcome Scale. *European Journal of Medical Genetics*, *62*(5), 324–334. https://doi.org/10.1016/j.ejmg.2018.11.015

Greany, T., & Earley, P. (2021). Introduction: School leadership and education system reform. In T. Greany & P. Earley (Eds.), *School leadership and education system reform* (2nd ed.). Bloomsbury Publishing.

Greiff, S., & Kyllonen, P. (2016a). Contemporary Assessment Challenges: The Measurement of 21st Century Skills. *Applied Measurement in Education*, *29*(4), 243–244. https://doi.org/10.1080/08957347.2016.1209209

Greiff, S., & Kyllonen, P. (2016b). Contemporary Assessment Challenges: The Measurement of 21st Century Skills. *Applied Measurement in Education*, *29*(4), 243–244. https://doi.org/10.1080/08957347.2016.1209209

Grek, S. (2009). Governing by numbers: The PISA ‘effect’ in Europe. *Journal of Education Policy*, *24*(1), 23–37. https://doi.org/10.1080/02680930802412669

Grek, S. (2010). International Organisations and the Shared Construction of Policy ‘Problems’: Problematisation and Change in Education Governance in Europe. *European Educational Research Journal*, *9*(3), 396–406. https://doi.org/10.2304/eerj.2010.9.3.396

Griffin, P., & Care, E. (2014). *Assessment and teaching of 21st century skills: Methods and approach*. Springer.

Griffin, P., Care, E., & Harding, S.-M. (2015). Task Characteristics and Calibration. In P. Griffin & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 133–178). https://doi.org/10.1007/978-94-017-9395-7_7

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

Gu, X., & Cai, H. (2019). How a semantic diagram tool influences transaction costs during collaborative problem solving. *Journal of Computer Assisted Learning*, *35*(1), 23–33. https://doi.org/10.1111/jcal.12307

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

Hao, J., Liu, L., von Davier, A., & Kyllonen, P. (2017). Initial Steps Towards a Standardized Assessment for Collaborative Problem Solving (CPS): Practical Challenges and Strategies. In A. von Davier, M. Zhu, & P. Kyllonen (Eds.), *Innovative Assessment of Collaboration* (pp. 135–156). Springer International Publishing. https://doi.org/10.1007/978-3-319-33261-1_9

Harding, S.-M. E., & Griffin, P. (2016). Rasch Measurement of Collaborative Problem Solving in an Online Environment. *Journal of Applied Measurement*, *1*(17), 35–53.

Harding, S.-M. E., Griffin, P., Awwal, N., Alom, B. M., & Scoular, C. (2017). Measuring Collaborative Problem Solving Using Mathematics-Based Tasks. *AERA Open*, *3*(3), 1–19. https://doi.org/10.1177/2332858417728046

Hathcoat, J. D., Curtis, N. A., Sanders, C. B., & Liu, S. (2018). Validity, History of. In B. B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (pp. 1777–1780). SAGE Publications, Inc. https://doi.org/10.4135/9781506326139

Herborn, K., Mustafić, M., & Greiff, S. (2017). Mapping an Experiment-Based Assessment of Collaborative Behavior Onto Collaborative Problem Solving in PISA 2015: A Cluster Analysis Approach for Collaborator Profiles. *Journal of Educational Measurement*, *54*(1), 103–122. https://doi.org/10.1111/jedm.12135

Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? *Computers in Human Behavior*, *104*, 105624. https://doi.org/10.1016/j.chb.2018.07.035

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A Framework for Teachable Collaborative Problem Solving Skills. In P. Griffin & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 37–56). Springer Netherlands. https://doi.org/10.1007/978-94-017-9395-7_2

Hmelo-Silver, C. E., Chinn, C. A., Chan, C. K. K., & O'Donnell, A. (Eds.). (2013). *The international handbook of collaborative learning*. Routledge.

Hopfenbeck, T. N. (2016). The power of PISA – limitations and possibilities for educational research. *Assessment in Education: Principles, Policy & Practice*, *23*(4), 423–426. https://doi.org/10.1080/0969594X.2016.1247518

Hopfenbeck, T. N., Lenkeit, J., Masri, Y. E., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research*, *62*(3), 333–353. https://doi.org/10.1080/00313831.2016.1258726

Hopfenbeck, T. N., & Maul, A. (2011). Examining Evidence for the Validity of PISA Learning Strategy Scales Based on Student Response Processes.

International Journal of Testing, *11*(2), 95–121. https://doi.org/10.1080/15305058.2010.529977

Howe, K. R. (1985). Two Dogmas of Educational Research. *Educational Researcher*, *14*(8), 10–18. https://doi.org/10.3102/0013189X014008010

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

Janssen, J., Kirschner, F., Erkens, G., Kirschner, P. A., & Paas, F. (2010). Making the Black Box of Collaborative Learning Transparent: Combining Process-Oriented and Cognitive Load Approaches. *Educational Psychology Review*, *22*(2), 139–154. https://doi.org/10.1007/s10648-010-9131-x

Jerrim, J. (2016). PISA 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice*, *23*(4), 495–518. https://doi.org/10.1080/0969594X.2016.1147420

Jerrim, J. (2021). PISA 2018 in England, Northern Ireland, Scotland and Wales: Is the data really representative of all four corners of the UK? *Review of Education*, *9*(3), e3270. https://doi.org/10.1002/rev3.3270

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C., & McKeown, C. (2018). PISA 2015: How big is the 'mode effect' and what has been done about it? *Oxford Review of Education*, *44*(4), 476–493. https://doi.org/10.1080/03054985.2018.1430025

Jerrim, J., Oliver, M., & Sims, S. (2019). The relationship between inquiry-based teaching and students' achievement. New evidence from a longitudinal PISA study in England. *Learning and Instruction*, *61*, 35–44. https://doi.org/10.1016/j.learninstruc.2018.12.004

Jerrim, J., Parker, P., Choi, A., Chmielewski, A. K., Sälzer, C., & Shure, N. (2018). How Robust Are Cross-Country Comparisons of PISA Scores to the Scaling Model Used? *Educational Measurement: Issues and Practice*, *37*(4), 28–39. https://doi.org/10.1111/emip.12211

Jerrim, J., & Shure, N. (2016). *Achievement of 15-Year-Olds in England: PISA 2015 National Report*. UCL Institute of Education, Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/574925/PISA-2015_England_Report.pdf

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational Researcher*, *33*(7), 14–26.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a Definition of Mixed Methods Research. *Journal of Mixed Methods Research*, *1*(2), 112–133. https://doi.org/10.1177/1558689806298224

Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: Can comparative judgement help? *Educational Studies in Mathematics*, *89*(3), 337–355. https://doi.org/10.1007/s10649-015-9607-1

Jones, I., Swan, M., & Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and*

Mathematics Education, *13*(1), 151–177. https://doi.org/10.1007/s10763-013-9497-6

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (Vol. 4, pp. 17–64). American Council on Education/Praeger Publishers.

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., Groot, E. D., Gilbert, M. C., Musu, L., Kempler, T. M., & Kelly, K. L. (2007). Cognitive Processing of Self-Report Items in Educational Research: Do They Think What We Mean? *Educational Psychologist*, *42*(3), 139–151. https://doi.org/10.1080/00461520701416231

Kelley, T., Ebel, R., & Linacre, J. M. (2002). Item Discrimination Indices. *Rasch Measurement Transactions*, *16*(3), 883–884.

Kreiner, S., & Christensen, K. B. (2014). Analyses of Model Fit and Robustness. A New Look at the PISA Scaling Model Underlying Ranking of Countries According to Reading Literacy. *Psychometrika*, *79*(2), 210–231. https://doi.org/10.1007/s11336-013-9347-z

Krkovic, K., Wüstenberg, S., & Greiff, S. (2016). Assessing Collaborative Behavior in Students: An Experiment-Based Assessment Approach. *European Journal of Psychological Assessment*, *32*(1), 52–60. https://doi.org/10.1027/1015-5759/a000329

Kumpulainen, K., & Kaartinen, S. (2003). The Interpersonal Dynamics of Collaborative Reasoning in Peer Interactive Dyads. *The Journal of Experimental Education*, *71*(4), 333–370. https://doi.org/10.1080/00220970309602069

Kuo, B.-C., Liao, C.-H., Pai, K.-C., Shih, S.-C., Li, C.-H., & Mok, M. M. C. (2020). Computer-based collaborative problem-solving assessment in Taiwan. *Educational Psychology*, *40*(9), 1164–1185. https://doi.org/10.1080/01443410.2018.1549317

Lai, E. R. (2011). *Collaboration: A literature review*. Pearson. http://images.pearsonassessments.com/images/tmrs/Collaboration-Review.pdf

Lave, J. (1991). Situating Learning in Communities of Practice. In L. Resnick, J. Levine, & S. Teasley (Eds.), *Perspectives on Socially Shared Cognition* (pp. 63–83). American Psychological Association.

Leech, N. L., & Onwuegbuzie, A. J. (2009). A typology of mixed methods research designs. *Quality & Quantity*, *43*(2), 265–275. https://doi.org/10.1007/s11135-007-9105-3

Li, C.-H., & Liu, Z.-Y. (2017). Collaborative Problem-Solving Behavior of 15-Year-Old Taiwanese Students in Science Education. *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(10), 6677–6695. https://doi.org/10.12973/ejmste/78189

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

Lin, P.-C., Hou, H.-T., & Chang, K.-E. (2020). The development of a collaborative problem solving environment that integrates a scaffolding mind tool and simulation-based learning: An analysis of learners' performance and their cognitive process in discussion. *Interactive Learning Environments*, 1–18. https://doi.org/10.1080/10494820.2020.1719163

Linacre, J. M. (n.d.-a). *Fit diagnosis: Infit outfit mean-square standardized: Winsteps Help*. Retrieved 2 August 2022, from https://www.winsteps.com/winman/misfitdiagnosis.htm

Linacre, J. M. (n.d.-b). *Table 23.99 Largest residual correlations for items: Winsteps Help*. Retrieved 2 August 2022, from https://www.winsteps.com/winman/table23_99.htm

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*.

Linacre, J. M. (2006a). *Dimensionality: Contrasts & variances*. https://www.winsteps.com/winman/principalcomponents.htm

Linacre, J. M. (2006b). *Reliability and separation of measures*. https://www.winsteps.com/winman/reliability.htm

Linacre, J. M. (2006c). *WINSTEPS Rasch measurement software*. Mesa Press.

Liu, O. L., Wilson, M., & Paek, I. (2008). A Multidimensional Rasch Analysis of Gender Differences in PISA Mathematics. *Journal of Applied Measurement*, *9*(1), 18–35.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Masters, G. N. (1999). Partial credit model. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 98–109). Elsevier Science.

Masters, G. N., & Keeves, J. P. (Eds.). (1999). *Advances in measurement in educational research and assessment* (1st ed). Pergamon.

Maul, A. (2018). Validity. In B. B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (pp. 1771–1775). SAGE Publications, Inc. https://doi.org/10.4135/9781506326139

Mayer, R. E. (1992). *Thinking, problem solving, cognition* (2nd ed.). Freeman.

Mayer, R. E. (2009). Information Processing. In T. L. Good (Ed.), *21st Century Education: A Reference Handbook* (pp. 168–174). SAGE Publications, Inc. https://doi.org/10.4135/9781412964012

Mayer, R. E. (2013). Problem Solving. In D. Reisberg (Ed.), *The Oxford Handbook of Cognitive Psychology* (pp. 769–778). Oxford University Press.

Mayer, R. E., & Wittrock, M. C. (2006). Problem Solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology* (2nd ed., pp. 287–304). Erlbaum.

Messick, S. (1980). Test Validity and the Ethics of Assessment. *American Psychologist*, *35*, 1012–1027.

Messick, S. (1987). *Validity* (pp. 1–209). Educational Testing Service. https://onlinelibrary.wiley.com/doi/10.1002/j.2330-8516.1987.tb00244.x

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.

Messick, S. (1995). Validity of Psychological Assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.

Meyer, H.-D. (2014). The OECD as Pivot of the Emerging Global Educational Accountability Regime: How Accountable are the Accountants? *Teachers College Record*, *116*(9), 1–20. https://doi.org/10.1177/016146811411600907

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus Article: On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research & Perspective*, *1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Annals of Internal Medicine*, *151*(4), 264–269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

Morgan, D. L. (2007). Paradigms Lost and Pragmatism Regained: Methodological Implications of Combining Qualitative and Quantitative Methods. *Journal of Mixed Methods Research*, *1*(1), 48–76. https://doi.org/10.1177/2345678906292462

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–177. https://doi.org/10.1177/014662169201600206

National Research Council. (2011). *Assessing 21st Century Skills: Summary of a Workshop*. National Academies Press (US). http://www.ncbi.nlm.nih.gov/books/NBK84218/

Nelson, L. M. (1999). Collaborative problem solving. In *Instructional design theories and models: A new paradigm of instructional theory* (Vol. 1–2, pp. 241–267). Lawrence Erlbaum Associates.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hall.

Nieminen, J. H., Chan, M. C. E., & Clarke, D. (2022). What affordances do open-ended real-life tasks offer for sharing student agency in collaborative problem-solving? *Educational Studies in Mathematics*, *109*(1), 115–136. https://doi.org/10.1007/s10649-021-10074-9

Nouri, J., Åkerfeldt, A., Fors, U., & Selander, S. (2017). Assessing Collaborative Problem Solving Skills in Technology-Enhanced Learning Environments – The PISA Framework and Modes of Communication. *International Journal of*

Emerging Technologies in Learning (IJET), *12*(04), 163. https://doi.org/10.3991/ijet.v12i04.6737

O'Donnell, A., & Hmelo-Silver, C. E. (2013). Introduction: What is Collaborative Learning? : An Overview. In *The international handbook of collaborative learning* (pp. 1–15). Routledge. https://doi.org/10.4324/9780203837290-1

OECD. (n.d.). *Description of the Released Unit from the 2015 PISA Collaborative Problem-Solving Assessment, Collaborative Problem-Solving Skills, and Proficiency Levels*. Retrieved 29 January 2022, from https://www.oecd.org/pisa/test/CPS-Xandar-scoring-guide.pdf

OECD. (2009). *PISA Data Analysis Manual: SPSS, Second Edition*. OECD. https://doi.org/10.1787/9789264056275-en

OECD. (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. OECD. https://doi.org/10.1787/9789264190511-en

OECD. (2017a). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. OECD Publishing. https://doi.org/10.1787/9789264281820-en

OECD. (2017b). *PISA 2015 Results (Volume V): Collaborative problem solving*. OECD Publishing. https://doi.org/10.1787/9789264285521-en

OECD. (2017c). *PISA 2015 Technical Report*. OECD Publishing. https://doi.org/10.1787/9789264273856-19-en

Oliveri, M. E., Lawless, R., & Molloy, H. (2017). A Literature Review on Collaborative Problem Solving for College and Workforce Readiness. *ETS Research Report Series*, *2017*(1), 1–27. https://doi.org/10.1002/ets2.12133

O'Neil, H. F. (1999). Perspectives on computer-based performance assessment of problem solving. *Computers in Human Behavior*, *15*(3), 255–268. https://doi.org/10.1016/S0747-5632(99)00022-9

O'Neil, H. F., Chuang, S.-H. (sabrina), & Chung, G. K. W. K. (2003). Issues in the Computer-based Assessment of Collaborative Problem Solving. *Assessment in Education: Principles, Policy & Practice*, *10*(3), 361–373. https://doi.org/10.1080/0969594032000148190

Ong, Y. M., Williams, J., & Lamprianou, I. (2013). Exploring differential bundle functioning in mathematics by gender: The effect of hierarchical modelling. *International Journal of Research & Method in Education*, *36*(1), 82–100. https://doi.org/10.1080/1743727X.2012.675263

Ong, Y. M., Williams, J. S., & Lamprianou, I. (2011). Exploration of the validity of gender differences in mathematics assessment using differential bundle functioning. *International Journal of Testing*, *11*(3), 271–293. https://doi.org/10.1080/15305058.2011.555574

Pampaka, M. (2021). Establishing Measurement Invariance across Time within an Accelerated Longitudinal Design. In A. Cernat & J. W. Sakshaug (Eds.), *Measurement Error in Longitudinal Data* (pp. 405–446). Oxford University Press. https://doi.org/10.1093/oso/9780198859987.003.0017

Pampaka, M., Williams, J., Hutcheson, G., Black, L., Davis, P., Hernandez-Martinez, P., & Wake, G. (2013). Measuring Alternative Learning Outcomes: Dispositions to study in Higher Education. *Journal of Applied Measurement*, *14*(2), 197–218.

Panayides, P., Robinson, C., & Tymms, P. (2010). The assessment revolution that has passed England by: Rasch measurement. *British Educational Research Journal*, *36*(4), 611–626. https://doi.org/10.1080/01411920903018182

Panayides, P., Robinson, C., & Tymms, P. (2015). Rasch measurement: A response to Goldstein. *British Educational Research Journal*, *41*(1), 180–182. https://doi.org/10.1002/berj.3182

Pepper, D., Hodgen, J., Lamesoo, K., Kõiv, P., & Tolboom, J. (2018). Think aloud: Using cognitive interviewing to validate the PISA assessment of student self-efficacy in mathematics. *International Journal of Research & Method in Education*, *41*(1), 3–16. https://doi.org/10.1080/1743727X.2016.1238891

Peterson, C. H., Peterson, N. A., & Powell, K. G. (2017). Cognitive Interviewing for Item Development: Validity Evidence Based on Content and Response Processes. *Measurement and Evaluation in Counseling and Development*, *50*(4), 217–223. https://doi.org/10.1080/07481756.2017.1339564

Petridou, A., & Williams, J. (2010a). The Extent of Mismeasurement for Aberrant Examinees. *Educational Assessment*, *15*(1), 42–68. https://doi.org/10.1080/10627191003673240

Petridou, A., & Williams, J. (2010b). Accounting for unexpected test responses through examinees' and their teachers' explanations. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 357–382. https://doi.org/10.1080/0969594X.2010.516606

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell.

Pólya, G. (1945). *How to solve it: A new aspect of mathematical method*. Princeton University Press.

Polyak, S. T., von Davier, A., & Peterschmidt, K. (2017). Computational Psychometrics for the Measurement of Collaborative Problem Solving Skills. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.02029

Prevett, P. S., Pampaka, M., Farnsworth, V. L., Kalambouka, A., & Shi, X. (2020). A Situated Learning Approach to Measuring Financial Literacy Self-Efficacy of Youth. *Journal of Financial Counseling and Planning*, *31*(2), 229–250. https://doi.org/10.1891/JFCP-18-00038

QSR International Pty Ltd. (2018). *NVivo* (Version 12) [Computer software]. https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

Ramalingam, D. (2016). *Using data from computer-delivered assessments to improve construct validity and measurement precision* [Melbourne Graduate School of Education]. http://hdl.handle.net/11343/197541

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

Rios, J. A., Ling, G., Pugh, R., Becker, D., & Bacall, A. (2020). Identifying Critical 21st-Century Skills for Workplace Success: A Content Analysis of Job Advertisements. *Educational Researcher*, *49*(2), 80–89. https://doi.org/10.3102/0013189X19890600

Rizvi, F. A., & Lingard, B. (2006). Globalization and the changing nature of the OECD's educational work. In B. Lingard, H. Lauder, P. Brown, J. Dillabough, & A. Halsey (Eds.), *Education Globalization and Social Change* (pp. 247–260). Oxford University Press. http://minerva-access.unimelb.edu.au/handle/11343/31469

Roschelle, J., & Teasley, S. D. (1995). The Construction of Shared Knowledge in Collaborative Problem Solving. In *Computer Supported Collaborative Learning* (pp. 69–97). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-85098-1_5

Rosen, Y. (2014). Comparability of Conflict Opportunities in Human-to-Human and Human-to-Agent Online Collaborative Problem Solving. *Technology, Knowledge and Learning*, *19*, 147–164. https://doi.org/10.1007/s10758-014-9229-1

Rosen, Y. (2015). Computer-based Assessment of Collaborative Problem Solving: Exploring the Feasibility of Human-to-Agent Approach. *International Journal of Artificial Intelligence in Education*, *25*(3), 380–406. https://doi.org/10.1007/s40593-015-0042-3

Rosen, Y., & Foltz, P. (2014). Assessing collaborative problem solving through automated technologies. *Research and Practice in Technology Enhanced Learning*, *9*, 389–410.

Roseth, C. J., Johnson, D. W., & Johnson, R. T. (2008). Promoting early adolescents' achievement and peer relationships: The effects of cooperative, competitive, and individualistic goal structures. *Psychological Bulletin*, *134*(2), 223–246. https://doi.org/10.1037/0033-2909.134.2.223

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Ruddock, G., Clausen-May, T., Purple, C., & Ager, R. (2006). *Validation study of the PISA 2000, PISA 2003 and TIMSS-2003 international studies of pupil attainment* [DfES Research Report 772]. National Foundation for Educational Research. https://dera.ioe.ac.uk/6448/1/RR772.pdf

Salas, E., Reyes, D., & Woods, A. (2017). The Assessment of Team Performance: Observations and Needs. In A. von Davier, M. Zhu, & P. Kyllonen (Eds.), *Innovative Assessment of Collaboration* (pp. 21–36). Springer International Publishing. https://doi.org/10.1007/978-3-319-33261-1_2

Sammons, P., & Davis, S. (2017). Mixed Methods Approaches and their Application in Educational Research. In *The BERA/SAGE Handbook of Educational Research: Two Volume Set* (Vol. 1–2, pp. 477–504). SAGE Publications Ltd. https://doi.org/10.4135/9781473983953

Schoenfeld, A. H. (1985). *Mathematical problem solving*. Academic Press.

Schoenfeld, A. H. (1987). Pólya, Problem Solving, and Education. *Mathematics Magazine*, *60*(5), 283–291. https://doi.org/10.1080/0025570X.1987.11977325

Scoular, C., & Care, E. (2019). A Generalized Scoring Process to Measure Collaborative Problem Solving in Online Environments. *Educational Assessment*, *24*(3), 213–234. https://doi.org/10.1080/10627197.2019.1615372

Scoular, C., & Care, E. (2020). Monitoring patterns of social and cognitive student behaviors in online collaborative problem solving assessments. *Computers in Human Behavior*, 105874. https://doi.org/10.1016/j.chb.2019.01.007

Scoular, C., Care, E., & Hesse, F. W. (2017). Designs for Operationalizing Collaborative Problem Solving for Automated Assessment. *Journal of Educational Measurement*, *54*(1), 12–35. https://doi.org/10.1111/jedm.12130

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

Scoular, C., Eleftheriadou, S., Ramalingam, D., & Cloney, D. (2020). Comparative analysis of student performance in collaborative problem solving: What does it tell us? *Australian Journal of Education*. https://doi.org/10.1177/0004944120957390

Seidouvy, A., & Schindler, M. (2019). An inferentialist account of students' collaboration in mathematics education. *Mathematics Education Research Journal*. https://doi.org/10.1007/s13394-019-00267-0

Shaw, S., & Child, S. (2017). Utilising technology in the assessment of collaboration: A critique of PISA's collaborative problem-solving tasks. *Research Matters: A Cambridge Assessment Publication*, *24*, 17–22.

Shaw, S., & Crisp, V. (2011). Tracing the evolution of validity in educational measurement: Past issues and contemporary challenges. *Research Matters: A Cambridge Assessment Publication*, *11*, 14–19.

Shiel, G., & Eivers, E. (2009). International comparisons of reading literacy: What can they tell us? *Cambridge Journal of Education*, *39*(3), 345–360. https://doi.org/10.1080/03057640903103736

Siddiq, F., & Scherer, R. (2017). Revealing the processes of students' interaction with a novel collaborative problem solving task: An in-depth analysis of think-aloud protocols. *Computers in Human Behavior*, *76*, 509–525. https://doi.org/10.1016/j.chb.2017.08.007

Slavin, R. E. (1980). Cooperative Learning. *Review of Educational Research*, *50*(2), 315–342. https://doi.org/10.3102/00346543050002315

Slof, B., Erkens, G., Kirschner, P. A., Janssen, J., & Jaspers, J. G. M. (2012). Successfully carrying out complex learning-tasks through guiding teams' qualitative and quantitative reasoning. *Instructional Science*, *40*(3), 623–643. https://doi.org/10.1007/s11251-011-9185-2

Smith, E. V. (2002). Understanding Rasch Measurement: Detecting and Evaluating the Impact of Multidimensionality Using Item Fit Statistics and Principal Components Analysis of Residuals. *Journal of Applied Measurement*, *3*(2), 205–231.

Stadler, M., Herborn, K., Mustafić, M., & Greiff, S. (2019). Computer-Based Collaborative Problem Solving in PISA 2015 and the Role of Personality. *Journal of Intelligence*, *7*(3), 15. https://doi.org/10.3390/jintelligence7030015

Stadler, M., Herborn, K., Mustafić, M., & Greiff, S. (2020a). The assessment of collaborative problem solving in PISA 2015: An investigation of the validity of the PISA 2015 CPS tasks. *Computers & Education*, *157*, 103964. https://doi.org/10.1016/j.compedu.2020.103964

Stadler, M., Herborn, K., Mustafić, M., & Greiff, S. (2020b). The assessment of collaborative problem solving in PISA 2015: An investigation of the validity of the PISA 2015 CPS tasks. *Computers & Education*, *157*, 103964. https://doi.org/10.1016/j.compedu.2020.103964

Stadler, M., Shubeck, K. T., Greiff, S., & Graesser, A. C. (2020). Some critical reflections on the special issue: Collaboration in the 21st century: The theory, assessment, and teaching of collaborative problem solving. *Computers in Human Behavior*, *104*, 106135. https://doi.org/10.1016/j.chb.2019.09.011

Stahl, G. (2006). *Group Cognition: Computer Support for Building Collaborative Knowledge*. The MIT Press.

Stahl, G. (2013). Theories of Cognition in Collaborative Learning. In C. E. Hmelo-Silver, C. A. Chinn, C. Chan, & A. M. O'Donnell (Eds.), *The International Handbook of Collaborative Learning*. Routledge. https://doi.org/10.4324/9780203837290.ch4

StataCorp. (2019). *Stata Statistical Software: Release 16*. StataCorp LLC.

Strauss, A., & Corbin, J. M. (1998). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications.

Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education*, *143*, 103672. https://doi.org/10.1016/j.compedu.2019.103672

Tallentire, J., & Shervin, J. (2022, June 9). Raised by Wolves – Will we always know the difference between AI and real people? *Science and Engineering*. https://www.mub.eps.manchester.ac.uk/science-engineering/2022/06/09/raised-by-wolves-will-we-always-know-the-difference-between-ai-and-real-people/

Tang, P., Liu, H., & Wen, H. (2021). Factors Predicting Collaborative Problem Solving: Based on the Data From PISA 2015. *Frontiers in Education*, *6*. https://www.frontiersin.org/articles/10.3389/feduc.2021.619450

Tashakkori, A., & Creswell, J. W. (2007). Editorial: The New Era of Mixed Methods. *Journal of Mixed Methods Research*, *1*(1), 3–7. https://doi.org/10.1177/2345678906293042

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. SAGE.

Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social & behavioral research*. SAGE Publications.

Tashakkori, A., & Teddlie, C. (Eds.). (2010). *Handbook of mixed methods in social & behavioral research* (2nd ed.). SAGE Publications.

Teddlie, C., & Sammons, P. (2010). Applications of mixed methods to the field of Educational Effectiveness Research. In B. P. M. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological Advances in Educational Effectiveness Research* (pp. 115–152). Routledge.

Teddlie, C., & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools*, *13*(1), 12–28.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. SAGE.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Lawrence Erlbaum Associates, Inc. https://doi.org/10.1075/z.62.13kok

Thomas, J., Brunton, J., & Graziosi, S. (2010). *EPPI-Reviewer 4: Software for research synthesis*. Social Science Research Unit, UCL Institute of Education.

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, *8*(1), 45. https://doi.org/10.1186/1471-2288-8-45

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. Macmillan Press. https://doi.org/10.5962/bhl.title.55072

Tourangeau, R. (1984). Cognitive science and survey methods: Acognitive perspective. In T. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey design: Building a bridge between disciplines* (pp. 73–100). National Academy Press.

Turing, A. M. (1950). I.—Computing machinery and intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Turner, R., & Adams, R. J. (2007). The Programme for International Student Assessment: An Overview. *Journal of Applied Measurement*, *8*(3), 237–248.

Vista, A., Care, E., & Awwal, N. (2017). Visualising and examining sequential actions as behavioural paths that can be interpreted as markers of complex behaviours. *Computers in Human Behavior*, *76*, 656–671. https://doi.org/10.1016/j.chb.2017.01.027

von Davier, A. (2017). Computational Psychometrics in Support of Collaborative Educational Assessments: Computational Psychometrics. *Journal of Educational Measurement*, *54*(1), 3–11. https://doi.org/10.1111/jedm.12129

von Davier, A., & Halpin, P. (2013). Collaborative Problem Solving and the Assessment of Cognitive Skills: Psychometric Considerations. *ETS Research Report Series*, *2013*(2), i–36. https://doi.org/10.1002/j.2333-8504.2013.tb02348.x

von Davier, A., Zhu, M., & Kyllonen, P. (Eds.). (2017). *Innovative Assessment of Collaboration*. Springer International Publishing. https://www.springer.com/gb/book/9783319332598

Vygotsky, L. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Webb, M., & Gibson, D. (2015). Technology enhanced assessment in complex collaborative settings. *Education and Information Technologies*, *20*(4), 675–695. https://doi.org/10.1007/s10639-015-9413-5

Whelehan, P., Pampaka, M., Boyd, J., Armstrong, S., Evans, A., & Ozakinci, G. (2021). Application of the Rasch measurement framework to mammography positioning data. *Data in Brief*, *38*, 107387. https://doi.org/10.1016/j.dib.2021.107387

Willis, G. B. (2005). *Cognitive interviewing a tool for improving questionnaire design*. SAGE.

Wolfe, E. W., & Smith, E. V. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I - instrument development tools. *Journal of Applied Measurement*, *8*(1), 97–123.

Wolfe, E. W., & Smith, E. V. (2007b). Instrument development tools and activities for measure validation using rasch models: Part II - Validation activities. *Journal of Applied Measurement*, *8*(2), 204–234.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every educator and psychologist should know* (pp. 65–104). Lawrence Erlbaum Associates.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, *70*(12), 857–860.

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

Wright, B. D., & Mok, M. (2000). Understanding Rasch measurement: Rasch models overview. *Journal of Applied Measurement*, *1*(1), 83–106.

Wright, B. D., & Mok, M. M. C. (2004). An overview of the family of rasch measurement models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to* *rasch measurement: Theory, models and applications* (pp. 1–24). Maple Grove. https://repository.eduhk.hk/en/publications/an-overview-of-the-family-of-rasch-measurement-models-4

Wright, B. D., & Stone, M. (1999). *Measurement Essentials* (2nd ed.). Wide Range, Inc.

Wu, M., & Adams, R. J. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Educational Measurement Solutions.

Yuan, J., Xiao, Y., & Liu, H. (2019). Assessment of Collaborative Problem Solving Based on Process Stream Data: A New Paradigm for Extracting Indicators and Modeling Dyad Data. *Frontiers in Psychology*, *10*, 369. https://doi.org/10.3389/fpsyg.2019.00369

Zhang, S., Cao, Y., Chan, M. C. E., & Wan, M. E. V. (2022). A comparison of meaning negotiation during collaborative problem solving in mathematics between students in China and Australia. *ZDM – Mathematics Education*, *54*(2), 287–302. https://doi.org/10.1007/s11858-022-01335-9

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, *2012*(1), i–30. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x

Chapter 10 Appendices

10.1 Appendix 1: University of Manchester guidance on journal-format thesis submission

Journal Format PhD Theses - Guiding Principles for Students and Staff

This information is provided as supplementary guidance to the main University 'Presentation of Theses' Policy which students should consult before starting to write the thesis. Students can submit their research in the traditional thesis format or the Journal Format and should ensure they read all available guidance before making the final decision on thesis format. In some disciplines the Journal Format thesis is the standard and it is therefore important that students also refer to any discipline specific guidance which is available via the relevant supervisor/ School/ Faculty.

1. INTRODUCTION TO THE JOURNAL FORMAT

i. Journal Format was formerly referred to as Alternative Format.

ii. The Journal Format thesis allows students to write sections of their doctoral thesis in a format suitable for publication in a peer-reviewed journal.

iii. One of the major considerations for submitting in Journal Format is the level of contribution that the student has made to the papers to be included in the thesis. The level of the student's contribution must be made explicitly clear within the thesis.

iv. Papers within the journal format do not have to be already published or even submitted for publication. It is recognised that eventual publications may differ from the chapters in the Journal Format thesis due to feedback from publishers, further research or developments in the subject.

v. Not all research projects will produce material suitable to present in Journal Format and consideration should be given to the most appropriate format for the research.

vi. The thesis should adhere to the basic principles of a traditional thesis, i.e. it must still represent an original contribution to the field of research, demonstrate an understanding of the entire body of work in the thesis, outline the relationship with existing literature and future developments and it must be a coherent body of related work.

vii. If it is appropriate to do so, and the main supervisor is in agreement, it may be possible to submit an MPhil / MD / Professional Doctorate/ Practice – Based/ DBA thesis in Journal Format.

2. ADVANTAGES OF THE JOURNAL FORMAT

i. Presenting research in the form of papers will help students to develop their skills in writing scholarly papers or other research outputs. These skills will be essential for a career as a researcher. (Note: it is important to recognise that the traditional format thesis will also develop writing skills and consideration should be given to the best approach for the student's research outcomes and discipline area).

ii. Sections of the paper (e.g. the method section or the results section) can be written-up and prepared as the student progresses through their programme. This avoids having to rewrite parts of the thesis to submit for publication at a later date.

iii. This format reduces the potential conflict of interest between the drive to publish papers and timely completion of the thesis as both can be achieved simultaneously.

iv. Encourages faster publication and enhances the student's research profile /career prospects.

3. CHALLENGES OF THE JOURNAL FORMAT

i. Not all Examiners are familiar with the Journal Format thesis. However, a thesis submitted

ii. in Journal Format is assessed on the same basis as any thesis and guidance is provided to Examiners on this type of thesis submission.

iii. It can be difficult for examiners to determine the student's individual contribution to publications. However, if the guidance is followed and students clearly document their own contribution throughout the thesis these problems should not arise.

4. DECIDING TO WRITE A JOURNAL FORMAT THESIS

i. The decision to submit in Journal Format will be part of the planning of the research project and students should discuss the format of their thesis with their main supervisor early on in their programme because this will give time to plan the structure and content of the thesis and also to set aside time to write the papers.

ii. Students will also be asked to comment on these discussions regarding thesis format as part of the annual expectations form at the start of each year.

iii. There is a potential conflict between producing multiple papers for the Journal Format thesis and producing 1 high impact paper (1 paper would not be sufficient for Journal Format). In some cases this decision can only be made once the results have been identified but this should be discussed between the student and the main supervisor prior to making any final decision on thesis format. In addition, it is worth noting that a high impact paper can be created from a fusion of thesis chapters formed of smaller papers.

iv. Depending on how the research develops and the analysis of data, there is flexibility on when students have to make the final decision regarding the type of thesis format submission and it may not be until Year 2 or 3 that students feel in a position to use the Journal Format. As with all aspects of the programme, planning the best approach for students, in conjunction with the main supervisor, will be the most effective way to manage the Journal Format.

v. If a decision is made to submit the thesis in Journal Format, the student should discuss their intention with their supervisor. The student should then declare their intention on thesis format on the Notice of Submission form.

vi. If students subsequently decide that a traditional thesis format might be more appropriate, they should carefully consider this course of action in terms of the time it takes to put the thesis together and discuss with their supervisor. Supervisors may not agree if it is felt that the student may not be able to submit on time. It is important that students do not leave this decision too late in the process.

5. STRUCTURE/CONTENT OF JOURNAL FORMAT THESES

i. Examples of other theses that have been successfully submitted in Journal Format (previously called Alternative Format) should be available via the supervisory team, the Graduate Office or via the institutional repository. It may not be possible to find an example in the exact research area so students are advised to review a few examples of successful submissions in the first instance. Full guidance on the format and structure required is provided in the Presentation of Theses policy. The thesis should include a general introduction and literature review to set the context and hypotheses. This should also include the details of each paper contained within the thesis and ideally a narrative of how these papers constitute a coherent body of work and relate to each other. It is particularly important for the Journal Format thesis that the aims and objectives are written to emphasise how the body of work interconnects. Furthermore, students can also include chapters on methodology and their critical evaluation of their studies, including a more detailed discussion and critique, than allowed for in a journal paper.

ii. The majority of results chapters should be presented as a ‘paper’ with an abstract, introduction, materials and methods, results, discussion and references.

iii. A final concluding discussion chapter (which should not be a repetition of previous chapters) should bring the thesis together and provide a critical evaluation of the findings, justify decisions made and set out ideas for future work. If not contained sufficiently within the papers, students may want to include supplementary information such as statistical data. The main supervisor can advise on relevant information to include.

iv. The number of papers included in the Journal Format thesis may vary according to discipline and is not prescribed, but should reflect the quantity, quality and originality of research and analysis expected of a candidate submitting a traditional thesis. There is no upper limit, but three to five papers or equivalent results chapters is typical. Students should also speak to their Faculty/ School about any discipline – specific guidance and consult with their main supervisor for advice. Ultimately the examiners will judge whether the quantity and quality of the work, the critical analysis and originality of the research and the defense of the thesis in the Viva Voce, justifies the award of a PhD so this must be taken into consideration when writing the thesis.

v. Students should ensure that the thesis is not weakened by lack of continuity and reasoning between chapters or by the separation of figures from the text they refer to.

vi. It is recommended that separate versions of the paper be inserted and that the pagination sequence should flow throughout the thesis rather than inserting pre-prints. Ideally, to ease readability, figures/tables and accompanying legends should be included at the appropriate point in the text of the papers, and not at the end of the text as would be typical for a manuscript submitted for publication.

vii. The journal Format offers flexibility in that students may include papers that have already been published or submitted or draft papers that have not yet been submitted or are not yet suitable for publication. Chapters can include various kinds of data and results including reviews, preliminary studies, pilot data, trial designs and lab results. Students are not precluded from presenting ‘negative’ results as long as they form a coherent part of the thesis. It is important to note that journal chapters which have not yet been submitted for publication may subsequently change when submitted for publication following input from co-authors, journal editors or peer review. Therefore journal formatted chapters may form a stepping stone towards a subsequent publication and there may be a long lag from thesis submission to publication.

viii. It is expected that students will have taken the major role in ALL aspects of production of the papers including: planning and execution, data acquisition, analysis and writing the paper. Where students have collaborated or co-authored any papers, the level of contribution must be made explicitly clear in the introduction of the thesis. Where students include a published paper which includes content authored by themselves within their thesis they must make it clear that the paper has already been published in order to avoid issues with self plagiarism.

ix. If data is not contained in sufficient detail within the published papers and is important to the thesis, it should be included in the same way as supplementary material for the journal i.e. in journal style (e.g. statistical data or a more detailed description of methods). Space restrictions do not apply to a thesis in the same way as restrictions on published work (see section 9 in the Presentation of Theses Policy for information on word count restrictions). Examiners will still want to see evidence of the detailed thought processes that led to the research design (including experimental design) and conclusions that are presented.

x. It is possible to add information to papers which are already in press or published. If the content from a published paper is significantly revised students will need to reference the paper at all appropriate points, otherwise this could be considered as self-plagiarism. See also sections related to IP/ Copyright/ Plagiarism.

xi. As noted in point vi it is not recommended to include pre-prints in the thesis but instead students are advised to insert a version of the paper. Where this presents problems, students can bind off-prints straight into the thesis. However, students may wish to consider reformatting if they are much smaller than A4 (or different sizes), so to be consistent with the overall presentational style of the thesis.

xii. All figures and tables should be legible and appear as close to the relevant text in the thesis as possible; this applies to both published and non-published material that is included in the thesis. Sometimes images/figures in published papers need to be placed according to best space fit.

xiii. As each paper will have a self-contained list of references and individual style depending on the journal requirements, students will need to consider making minor formatting / stylistic adjustments so that the thesis has consistency (e.g.: references should all be provided in the same format). Papers should be presented in such a way as to assist the examiners' reading of the thesis in the best way possible. References associated with the introduction and concluding chapters should be presented in the most appropriate format. Students should consult their main supervisor and any discipline-specific guidance on this.

6. IP / COPYRIGHT / PLAGIARISM

i. Students will be required to sign a declaration that the thesis is their own work and that they have not submitted the work for another qualification. Students should explain and fully justify the nature and extent of their own contribution and the contribution of co-authors and other collaborators in the introductory part of the thesis and anywhere else appropriate throughout the body of the thesis. Students should consult University guidance on plagiarism for further guidance.

ii. It is advisable for students to discuss their stated contribution to each paper with their main supervisor and co- authors. Even if the student is the first author, the main supervisor or others may have contributed to the paper and the student needs to clarify the contribution of others. In some cases, it may be reasonable for a student to be asked to revise a paper chapter in order to reflect their own contribution more directly. Examiners will expect students to defend all of the work in any paper that forms part of the thesis, even if the work has been done (and acknowledged as such) by someone else.

iii. If appropriate, students can state their contribution in individual chapters relating to specific publications.

iv. Generally, unless IP has been signed over to a third party, and the student has solely created the IP and is not a member of staff, the student owns the IP they have created. However, it is expected that the student obtains permission from all co-authors for any paper that is included in the thesis. Most publishers request that students sign over copyright of any published material once published. Students should seek copyright permission from the publisher for any published work included in the thesis that isn't published in an 'open-access' journal. Where the publisher owns the copyright, permission from collaborators/co-authors would not therefore be needed.

v. Any concerns about IP should be discussed with the main supervisor in the first instance. UMIP also offer advice on IP and copyright regulations.

vi. Any sections which are copied from any published materials must be referenced appropriately, otherwise the student would be plagiarising material, even if the student was the original author of this material. If sections of the student's own papers are used without the appropriate references this will be considered as self-plagiarism.

7. EXAMINATION OF JOURNAL FORMAT THESES

i. The examination process will be exactly the same as for a traditional thesis. The examiner will be informed that the thesis has been presented as Journal Format and the School/Faculty office will provide them with the links to University guidance and policy documents on thesis submissions and Journal Format. It should be made clear to examiners that there will inevitably be some degree of repetition in the Journal Format thesis due to chapters being self-contained papers and background literature and issues being repeated. Students should not be penalised or asked to correct work on the basis of repetition within journal style chapters.

ii. A major consideration when preparing a Journal Format Thesis is that the examiners can follow and understand the thesis as a coherent body of work. Students should ensure that their thesis does not lack a full explanation of technical detail and consideration of controls because it is in the publication style format. The examiners will expect the thesis to demonstrate rigour in all aspects of the research. As noted earlier supplementary chapters containing methodological details such as raw data etc. may be included.

iii. The entire thesis is subject to scrutiny, **including** any peer-reviewed or published papers. The examiners are effectively another set of peer reviewers who are looking at the published papers in the context of the whole thesis. There are often examples where peer-reviewed work contains mistakes, errors or points of contention and so the student may still be required to correct, supplement, or explain all work presented for examination, even if it has already passed through a separate peer review as part of the publishing process.

iv. It is recommended that supervisors and/or internal examiners speak to the external examiner prior to submission to ensure they are aware of the requirements of submitting a thesis in Journal Format.

8. PUBLICATIONS AND OPEN ACCESS

i. Students should discuss their ‘publication strategy’ with their supervisors and check with their Faculty/ School for local discipline – specific guidance, from an early point in the programme. Students would need to consider the journals that they would target for publication of their papers and review their position on prior publication of work. Most publishers do not view work that has appeared in a thesis as ‘prior publication’ and in these scenarios, the thesis should be made open access, but the viewpoint of each publisher can vary. If in exceptional circumstances, a publisher does consider the thesis as prior publication, advice must be sought from the publisher to determine whether making the thesis open access would impact future publication of the work.

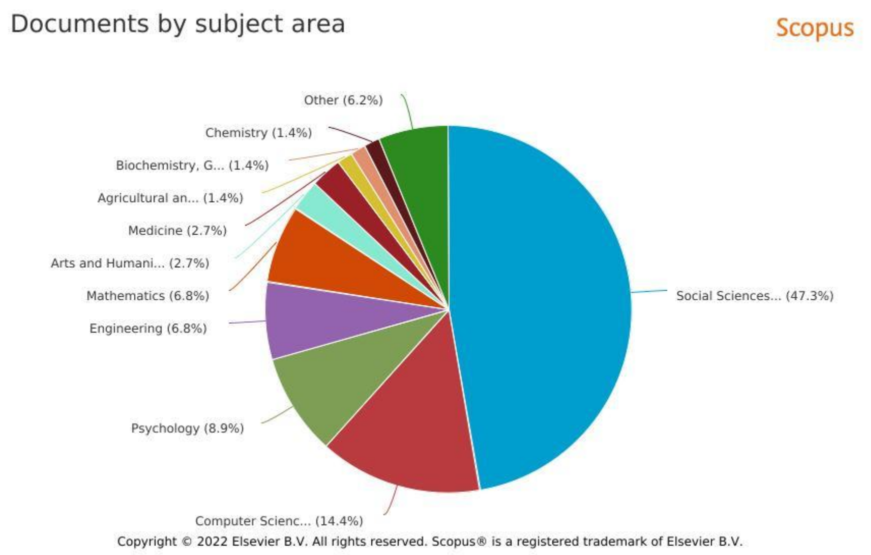
ii. The access setting on the theses may also be dependent on funder terms and conditions and students should check with their funder whether there are any contractual requirements.

10.2 Appendix 2: Updated literature search – July 2022

An updated literature search was run on the 21st of July 2022 in the Scopus database using the same keywords and restrictions to check for the number of new articles published between 2021 and 2022, even though these were not considered as part of the results (Table 10.1). The analysis of these was beyond the scope of this study, since the aim of the conceptual review was to inform the empirical phases following the research design. Therefore, integration with the results was not considered at this time, but future research work may consider extending the search. The search resulted in 85 articles published in this period, about half of them being indexed under the subject area Social Sciences, which includes Education as well as 20 more other subjects (Figure 10.1).

Table 10.1 Updated literature search in Scopus database

Date of search	Search string	Results
21 July 2022	(TITLE-ABS-KEY ("collaborative problem solving")) AND (TITLE-ABS-KEY (student* OR learner* OR pupil*)) AND (DOCTYPE (ar) OR DOCTYPE (re) OR PUBSTAGE (aip)) AND (LIMIT-TO (SRCTYPE , "j")) AND (LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2022)) AND (LIMIT-TO (LANGUAGE , "English"))	N = 85



Copyright © 2022 Elsevier B.V. All rights reserved. Scopus® is a registered trademark of Elsevier B.V.

Figure 10.1 Publications by subject area

A tentative list of potentially relevant papers, found in the updated search, is presented below to be included in future work reviewing the topic:

Andrews-Todd, J., Steinberg, J., Flor, M., & Forsyth, C. M. (2022). Exploring Automated Classification Approaches to Advance the Assessment of Collaborative Problem Solving Skills. *Journal of Intelligence*, *10*(3). https://doi.org/10.3390/jintelligence10030039

Chan, M. C. E., Moate, J., Clarke, D., Cunnington, R., Díez-Palomar, J., Friesen, M., Haataja, E., Hošpesová, A., Kuntze, S., Nieminen, J., Novotná, J., Ochoa, X., Sherwell, C., Tran, D., Tuohilamp, L., & with, the S. U. of L. project team. (2022). Learning research in a laboratory classroom: A reflection on complementarity and commensurability among multiple analytical accounts. *ZDM - Mathematics Education*, *54*(2), 317–329. https://doi.org/10.1007/s11858-022-01330-0

Gao, Q., Zhang, S., Cai, Z., Liu, K., Hui, N., & Tong, M. (2022). Understanding student teachers' collaborative problem solving competency: Insights from process data and multidimensional item response theory. *Thinking Skills and Creativity*, *45*. https://doi.org/10.1016/j.tsc.2022.101097

Kim, K., & Tawfik, A. A. (2021). Different approaches to collaborative problem solving between successful versus less successful problem solvers: Tracking changes of knowledge structure. *Journal of Research on Technology in Education*. https://doi.org/10.1080/15391523.2021.2014374

Nieminen, J. H., Chan, M. C. E., & Clarke, D. (2022). What affordances do open-ended real-life tasks offer for sharing student agency in collaborative problem-solving? *Educational Studies in Mathematics*, *109*(1), 115–136. https://doi.org/10.1007/s10649-021-10074-9

Tang, H., Dai, M., Yang, S., Du, X., Hung, J.-L., & Li, H. (2022). Using multimodal analytics to systemically investigate online collaborative problem-solving. *Distance Education*, *43*(2), 290–317. https://doi.org/10.1080/01587919.2022.2064824

Tang, P., Liu, H., & Wen, H. (2021). Factors Predicting Collaborative Problem Solving: Based on the Data From PISA 2015. *Frontiers in Education*, *6*. https://www.frontiersin.org/articles/10.3389/feduc.2021.619450

Unal, E., & Cakir, H. (2021). The effect of technology-supported collaborative problem solving method on students' achievement and engagement. *Education and Information Technologies*, *26*(4), 4127–4150. https://doi.org/10.1007/s10639-021-10463-w

Zhang, S., Cao, Y., Chan, M. C. E., & Wan, M. E. V. (2022). A comparison of meaning negotiation during collaborative problem solving in mathematics between students in China and Australia. *ZDM – Mathematics Education*, *54*(2), 287–302. https://doi.org/10.1007/s11858-022-01335-9

10.3 Appendix 3: PISA 2015 Test design

Table 10.2 presents the total of 66 different test forms created for PISA 2015 computer-based delivery and the respective sample sizes for England. Test forms that included at least one cluster with CPS items are highlighted with grey.

Table 10.2. PISA 2015 test design for computer-based assessment and sample size

Test form	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Frequencies N=students	%	Cumulative %
31	S	S	R1	R2	131	2.52	2.52
32	S	S	R2	R3	140	2.70	5.22
33	S	S	R3	R4	142	2.73	7.95
34	S	S	R4	R5	132	2.54	10.49
35	S	S	R5	R6	144	2.77	13.27
36	S	S	R6	R1	140	2.70	15.96
37	R1	R3	S	S	145	2.79	18.75
38	R2	R4	S	S	146	2.81	21.56
39	R3	R5	S	S	141	2.71	24.28
40	R4	R6	S	S	151	2.91	27.19
41	R5	R1	S	S	153	2.95	30.13
42	R6	R2	S	S	137	2.64	32.77
43	S	S	M1	M2	143	2.75	35.52
44	S	S	M2	M3	146	2.81	38.33
45	S	S	M3	M4	148	2.85	41.18
46	S	S	M4	M5	141	2.71	43.90
47	S	S	M5	M6	146	2.81	46.71
48	S	S	M6	M1	133	2.56	49.27
49	M1	M3	S	S	135	2.60	51.87
50	M2	M4	S	S	134	2.58	54.45
51	M3	M5	S	S	150	2.89	57.34
52	M4	M6	S	S	158	3.04	60.38
53	M5	M1	S	S	143	2.75	63.13
54	M6	M2	S	S	131	2.52	65.65
55	S	S	M1	R1	28	0.54	66.19
56	S	S	R2	M2	12	0.23	66.42
57	S	S	M3	R3	15	0.29	66.71
58	S	S	R4	M4	17	0.33	67.04
59	S	S	M5	R5	14	0.27	67.31
60	S	S	R6	M6	13	0.25	67.56
61	R1	M1	S	S	24	0.46	68.02

62	M2	R2	S	S	18	0.35	68.37
63	R3	M3	S	S	12	0.23	68.60
64	M4	R4	S	S	20	0.39	68.98
65	R5	M5	S	S	8	0.15	69.14
66	M6	R6	S	S	19	0.37	69.50
67	S	S	CPS1	M1	24	0.46	69.97
68	S	S	M2	CPS2	15	0.29	70.25
69	S	S	CPS3	M3	17	0.33	70.58
70	S	S	M4	CPS3	21	0.40	70.99
71	S	S	CPS2	M5	11	0.21	71.20
72	S	S	M6	CPS1	21	0.40	71.60
73	M1	CPS2	S	S	21	0.40	72.01
74	CPS3	M2	S	S	15	0.29	72.29
75	M3	CPS1	S	S	12	0.23	72.53
76	CPS1	M4	S	S	26	0.50	73.03
77	M5	CPS3	S	S	10	0.19	73.22
78	CPS2	M6	S	S	25	0.48	73.70
79	S	S	R1	CPS1	13	0.25	73.95
80	S	S	CPS2	R2	14	0.27	74.22
81	S	S	R3	CPS3	21	0.40	74.62
82	S	S	CPS3	R4	17	0.33	74.95
83	S	S	R5	CPS2	15	0.29	75.24
84	S	S	CPS1	R6	23	0.44	75.68
85	CPS2	R1	S	S	19	0.37	76.05
86	R2	CPS3	S	S	11	0.21	76.26
87	CPS1	R3	S	S	13	0.25	76.51
88	R4	CPS1	S	S	20	0.39	76.90
89	CPS3	R5	S	S	13	0.25	77.15
90	R6	CPS2	S	S	28	0.54	77.69
91	S	S	CPS1	CPS2	199	3.83	81.52
92	S	S	CPS2	CPS3	198	3.81	85.33
93	S	S	CPS3	CPS1	190	3.66	88.99
94	CPS2	CPS1	S	S	199	3.83	92.82
95	CPS3	CPS2	S	S	186	3.58	96.40
96	CPS1	CPS3	S	S	187	3.60	100.00
					Total: 5,194	Total: 100%	

Notes: Sample size for England in PISA 2015. S indicates science clusters, M1-M6 indicates mathematics clusters, R1-R6 indicates reading clusters, and CPS1-CPS3 indicates collaborative problem-solving clusters. Grey highlighting indicates test forms that included CPS cluster(s).

10.4 Appendix 4: PISA 2015 CPS items

Table 10.3 presents details about the scoring of all CPS items (n = 117) included in the computer-simulated, scenario-based assessment units of the PISA 2015 CPS assessment. The mapping of items to the CPS competencies and skills is based on information from the official PISA Results report (OECD, 2017b).

Table 10.3. Items included in the PISA 2015 CPS assessment (OECD, 2017b)

Item	Item code in PISA 2015	Name of CPS assessment unit	CPS competence	CPS Skill	Scoring
1	CC104101	Meeting in the Park	Taking actions	B2	0, 1
2	CC104102	Meeting in the Park	Shared understanding	A1	0, 1
3	CC104103	Meeting in the Park	Taking actions	A2	0, 1
4	CC104105	Meeting in the Park	Shared understanding	B1	0, 1
5	CC104106	Meeting in the Park	Shared understanding	B1	0, 1
6	CC104107	Meeting in the Park	Shared understanding	B1	0, 1
7	CC104201	Meeting in the Park	Team organisation	B3	0, 1
8	CC104202	Meeting in the Park	Shared understanding	B1	0, 1
9	CC104203	Meeting in the Park	Shared understanding	B1	0, 1
10	CC104204	Meeting in the Park	Team organisation	D3	0, 1
11	CC104205	Meeting in the Park	Taking actions	B2	0, 1
12	CC104206	Meeting in the Park	Shared understanding	C1	0, 1
13	CC104301C	Meeting in the Park	Shared understanding	C1	0, 1, 2, 3
14	CC104305	Meeting in the Park	Taking actions	D2	0, 1
15	CC104306	Meeting in the Park	Taking actions	C2	0, 1
16	CC106101	Making a Film	Shared understanding	A1	0, 1
17	CC106102	Making a Film	Shared understanding	A1	0, 1
18	CC106103	Making a Film	Shared understanding	A1	0, 1
19	CC106104	Making a Film	Team organisation	C3	0, 1
20	CC106105	Making a Film	Taking actions	A2	0, 1
21	CC106106	Making a Film	Shared understanding	A1	0, 1
22	CC106107C	Making a Film	Team organisation	C3	0, 1, 2, 3
23	CC106201	Making a Film	Shared understanding	A1	0, 1
24	CC106202	Making a Film	Team organisation	B3	0, 1
25	CC106203	Making a Film	Team organisation	B3	0, 1
26	CC106204	Making a Film	Taking actions	B2	0, 1
27	CC106205	Making a Film	Taking actions	B2	0, 1
28	CC106206	Making a Film	Team organisation	D3	0, 1
29	CC106207	Making a Film	Team organisation	C3	0, 1
30	CC106208	Making a Film	Team organisation	C3	0, 1
31	CC106209	Making a Film	Taking actions	C2	0, 1, 2
32	CC106301	Making a Film	Shared understanding	A1	0, 1
33	CC106302	Making a Film	Shared understanding	A1	0, 1

34	CC106303	Making a Film	Shared understanding	A1	0, 1
35	CC106304	Making a Film	Shared understanding	D1	0, 1
36	CC106305	Making a Film	Shared understanding	D1	0, 1
37	CC106306	Making a Film	Team organisation	C3	0, 1
38	CC106307	Making a Film	Taking actions	C2	0, 1, 2
39	CC102101	Field Trip	Team organisation	C3	0, 1
40	CC102102C	Field Trip	Shared understanding	D1	0, 1, 2
41	CC102201	Field Trip	Shared understanding	B1	0, 1
42	CC102202	Field Trip	Taking actions	D2	0, 1
43	CC102203	Field Trip	Team organisation	C3	0, 1
44	CC102204	Field Trip	Taking actions	B2	0, 1
45	CC102205	Field Trip	Team organisation	C3	0, 1
46	CC102206	Field Trip	Shared understanding	B1	0, 1
47	CC102207	Field Trip	Team organisation	C3	0, 1
48	CC102209C	Field Trip	Taking actions	C2	0, 1, 2, 3
49	CC102212	Field Trip	Taking actions	C2	0, 1
50	CC102213	Field Trip	Taking actions	C2	0, 1
51	CC103101	Preparing a Presentation	Shared understanding	A1	0, 1
52	CC103102	Preparing a Presentation	Shared understanding	A1	0, 1
53	CC103103	Preparing a Presentation	Shared understanding	A1	0, 1
54	CC103104	Preparing a Presentation	Shared understanding	A1	0, 1
55	CC103105	Preparing a Presentation	Shared understanding	A1	0, 1
56	CC103106	Preparing a Presentation	Shared understanding	A1	0, 1
57	CC103107	Preparing a Presentation	Shared understanding	A1	0, 1
58	CC103108C	Preparing a Presentation	Shared understanding	A1	0, 1, 2, 3, 4
59	CC103201	Preparing a Presentation	Team organisation	C3	0, 1
60	CC103202	Preparing a Presentation	Team organisation	C3	0, 1
61	CC103203	Preparing a Presentation	Shared understanding	B1	0, 1
62	CC103204	Preparing a Presentation	Shared understanding	D1	0, 1
63	CC103205	Preparing a Presentation	Shared understanding	D1	0, 1
64	CC103206	Preparing a Presentation	Shared understanding	B1	0, 1

65	CC103207	Preparing a Presentation	Shared understanding	D1	0, 1
66	CC103209	Preparing a Presentation	Taking actions	C2	0, 1
67	CC103210	Preparing a Presentation	Taking actions	C2	0, 1
68	CC103211	Preparing a Presentation	Taking actions	C2	0, 1
69	CC103301	Preparing a Presentation	Shared understanding	D1	0, 1
70	CC103302	Preparing a Presentation	Shared understanding	B1	0, 1
71	CC103303	Preparing a Presentation	Shared understanding	D1	0, 1
72	CC103304	Preparing a Presentation	Shared understanding	B1	0, 1
73	CC103305	Preparing a Presentation	Shared understanding	B1	0, 1
74	CC103306	Preparing a Presentation	Shared understanding	B1	0, 1
75	CC103307	Preparing a Presentation	Shared understanding	B1	0, 1
76	CC103308	Preparing a Presentation	Taking actions	C2	0, 1
77	CC103309	Preparing a Presentation	Taking actions	C2	0, 1
78	CC100101	Xandar	Team organisation	C3	0, 1
79	CC100102	Xandar	Shared understanding	C1	0, 1
80	CC100103	Xandar	Shared understanding	B1	0, 1
81	CC100104	Xandar	Shared understanding	B1	0, 1
82	CC100105	Xandar	Team organisation	B3	0, 1
83	CC100201	Xandar	Shared understanding	A1	0, 1
84	CC100202	Xandar	Team organisation	B3	0, 1
85	CC100203	Xandar	Team organisation	B3	0, 1
86	CC100301	Xandar	Team organisation	C3	0, 1
87	CC100302	Xandar	Shared understanding	D1	0, 1
88	CC100401	Xandar	Taking actions	D2	0, 1
89	CC100402	Xandar	Team organisation	D3	0, 1
90	CC105101	The Garden	Shared understanding	B1	0, 1, 2
91	CC105102	The Garden	Shared understanding	A1	0, 1
92	CC105103C	The Garden	Shared understanding	C1	0, 1, 2
93	CC105105C	The Garden	Taking actions	C2	0, 1, 2
94	CC105108C	The Garden	Taking actions	C2	0, 1, 2
95	CC105201C	The Garden	Shared understanding	A1	0, 1, 2
96	CC105203C	The Garden	Team organisation	B3	0, 1, 2, 3
97	CC105205	The Garden	Shared understanding	D1	0, 1, 2

98	CC105206	The Garden	Taking actions	C2	0, 1
99	CC105207	The Garden	Team organisation	D3	0, 1, 2
100	CC105208C	The Garden	Taking actions	C2	0, 1, 2, 3
101	CC105211	The Garden	Team organisation	D3	0, 1
102	CC105212C	The Garden	Taking actions	C2	0, 1, 2
103	CC105214	The Garden	Team organisation	D3	0, 1
104	CC105301	The Garden	Shared understanding	B1	0, 1
105	CC105302	The Garden	Shared understanding	B1	0, 1
106	CC105303	The Garden	Shared understanding	D1	0, 1
107	CC105304C	The Garden	Shared understanding	C1	0, 1
108	CC105306	The Garden	Shared understanding	B1	0, 1
109	CC105307	The Garden	Shared understanding	B1	0, 1, 2
110	CC105308C	The Garden	Team organisation	D3	0, 1, 2, 3
111	CC105401	The Garden	Shared understanding	D1	0, 1
112	CC105402	The Garden	Shared understanding	B1	0, 1
113	CC105403	The Garden	Shared understanding	B1	0, 1
114	CC105404	The Garden	Shared understanding	B1	0, 1
115	CC105406	The Garden	Team organisation	B3	0, 1
116	CC105407	The Garden	Team organisation	C3	0, 1
117	CC105408C	The Garden	Team organisation	D3	0, 1, 2

Notes: A1 = Discovering perspectives and abilities of team members, A2 = Discovering the type of collaborative interaction to solve the problem, along with goals, A3 = Understanding roles to solve the problem, B1 = Building a shared representation and negotiating the meaning of the problem, B2 = Identifying and describing tasks to be completed, B3 = Describing roles and team organisation, C1 = Communicating with team members about the actions to be/being performed, C2 = Enacting plans, C3 = Following rules of engagement, D1 = Monitoring and repairing the shared understanding, D2 = Monitoring results of actions and evaluating success in solving the problem, D3 = Monitoring, providing feedback and adapting the team organisation and roles

10.5 Appendix 5: Descriptive statistics for variables

Table 10.4. Descriptive statistics for continuous variables used in further statistical analysis

Variable	Mean	SD	Min	Max
CPS competence overall scale	0.48	0.82	-2.15	3.14
Sub-scale 1 (Shared understanding)	0.53	0.95	-2.30	4.50
Sub-scale 2 (Taking actions)	0.51	1.03	-4.71	4.44
Sub-scale 3 (Team organisation)	0.42	0.97	-2.63	4.28
Economic, social, and cultural status index	0.21	0.84	-2.49	2.91
Valuing teamwork	-0.07	0.96	-2.83	2.14
Valuing relationships	-0.06	0.96	-3.33	2.29
Science performance	510.68	99.20	208.21	834.40
Mathematics performance	491.95	95.78	112.50	796.61
Reading performance	497.74	97.33	148.79	810.85

Notes: Analytical sample n = 1,485

Table 10.5. Descriptive statistics for nominal variables used in further statistical analysis

Variable	Categories	N
Gender	0 = Male	771
	1 = Female	714
Geographic region	1 = Greater London	192
	2 = South	424
	3 = Midlands	450
	4 = North	419
School type	1 = Academy	683
	2 = Maintained selective	142
	3 = Maintained non-selective	528
	4 = Independent	132

Notes: Analytical sample n = 1,485

10.6 Appendix 6: Sample size and missing data

Table 10.6 reports on the variables used in statistical modelling and the size of missing data. The student sample for England after listwise exclusion of cases with missing data was n = 1,485.

Table 10.6. Variables used in further statistical analysis

Variable	Variable type	Sample size	Cases with missing data	% Over the total sample size
CPS competence overall scale	Continuous	1,584	2	0.13%
Sub-scale 1 (Shared understanding)	Continuous	1,584	2	0.13%
Sub-scale 2 (Taking actions)	Continuous	1,584	3	0.19%
Sub-scale 3 (Team organisation)	Continuous	1,584	3	0.19%
Economic, social, and cultural status index	Continuous	1,584	69	4.36%
Valuing teamwork	Continuous	1,584	59	3.72%
Valuing relationships	Continuous	1,584	52	3.28%
Gender	Nominal	1,584	0	0
Geographic region	Nominal	1,584	0	0
School type	Nominal	1,584	0	0
Science performance	Continuous	1,584	0	0
Mathematics performance	Continuous	1,584	0	0
Reading performance	Continuous	1,584	0	0

Notes: Student sample used is for England.

10.7 Appendix 7: Exploring multicollinearity issues

To check for potential multicollinearity issues, preliminary analysis included correlations between the independent variables.

Table 10.7. Correlation between independent variables

	ESCS	Valuing teamwork	Valuing relationships	Science performance	Mathematics performance	Reading performance
ESCS	1					
Valuing teamwork	-0.03	1				
Valuing relationships	0.10*	0.47*	1			
Science performance	0.32*	-0.16*	0.14*	1		
Mathematics performance	0.31*	-0.13*	0.12*	0.85*	1	
Reading performance	0.30*	-0.12*	0.17*	0.84*	0.74*	1

Notes: Asterisk indicates significance at p<0.05 level. Correlations above 0.5 shaded in grey.

10.8 Appendix 8: Cognitive interviewing – Anticipated probes

Table 10.8 presents the full list of anticipated probes developed for every Xandar item.

Table 10.8. Anticipated probes

Item	Verbal probe	Cognitive process targeted
X1	Was there anything confusing about this part?	Understanding
X2	Did it seem like one of the responses is supposed to be the right answer?	Judgement
	Was there an answer you wanted to give that was not available in the response options?	Response
	Were there response options that didn’t make sense to you?	Response
X3	Did it seem like one of the responses is supposed to be the right answer?	Judgement
	Was there an answer you wanted to give that was not available in the response options?	Response
	Were there response options that didn’t make sense to you?	Response
X4	Did it seem like one of the responses is supposed to be the right answer?	Judgement
	Was there an answer you wanted to give that was not available in the response options?	Response
	Were there response options that didn’t make sense to you?	Response
X5	Did it seem like one of the responses is supposed to be the right answer?	Judgement
	Was there an answer you wanted to give that was not available in the response options?	Response
	Were there response options that didn’t make sense to you?	Response
X6	Did this part feel awkward to answer? Why?	Judgement
	Were there response options that didn’t make sense to you?	Response
X7	Did this part feel awkward to answer? Why?	Judgement
	Did it seem like one of the responses is supposed to be the right answer?	Response
		Response

	Was there an answer you wanted to give that was not available in the response options?	Response
	Were there response options that didn’t make sense to you?	
X8	How comfortable did you feel answering this part?	Judgement
	Did it seem like one of the responses is supposed to be the right answer?	Response
	Was there an answer you wanted to give that was not available in the response options?	Response
X9	Was there anything confusing about this part?	Understanding
X10	How comfortable did you feel answering this part?	Judgement
	Did it seem like one of the responses is supposed to be the right answer?	Response
	Was there an answer you wanted to give that was not available in the response options?	Response
X11	Did it seem like one of the responses is supposed to be the right answer?	Judgement
	Was there an answer you wanted to give that was not available in the response options?	Response
	Were there response options that didn’t make sense to you?	Response
X12	Did it seem like one of the responses is supposed to be the right answer?	Judgement
	Was there an answer you wanted to give that was not available in the response options?	Response
	Were there response options that didn’t make sense to you?	Response

Notes. Verbal probes adapted from Peterson et al. (2017) and Willis (2005).

10.9 Appendix 9: Invitation to schools – Initial and Extended

MANCHESTER
1824
The University of Manchester

Dear [insert name of head teacher],

An Exploration of Student Responses to a Collaborative Problem Solving Task

We would like to invite your school to take part in a research study. In particular, Sofia Eleftheriadou, a research student from the University of Manchester, is conducting a study to find out more about young people's collaborative problem solving skills and we will be very grateful if you can help.

The study aims to find out how young people aged 14 to 16 years old respond to a collaborative problem solving task. We are asking students in Years 9 to 11 to complete an online task alone and as part of a group of two and discuss with the researcher the responses they have given and the way they approached the task. This will allow us to understand how young people answer a collaborative problem solving task. We are interested in particular in comparing students' answering the task alone and as part of a small group (dyad) because we believe that there might be a difference between the two.

We very much hope that this project is of interest to you and if you are interested we will appreciate it if you let us know which person will be best to liaise with so that we can send them more details about the project, the way in which your school and your students can help us and the administration of the task.

Finally, we are planning to organise an event about Collaborative problem solving where teachers from participating schools will be invited. In this event, we will disseminate some preliminary findings from a literature review that has been conducted, we will demonstrate a task assessing collaboration skills, and we will also disseminate some preliminary results.

We look forward to hearing from you as soon as possible.

With best wishes,
Sofia Eleftheriadou and Supervisors

Sofia Eleftheriadou | PhD Researcher | Manchester Institute of Education (MIE) |School of Environment, Education and Development | The University of Manchester | Email: sofia.eleftheriadou@manchester.ac.uk

Manchester Institute of
Education Ellen Wilkinson
Building
University of Manchester
M13 9PL

18 February 2019

An Exploration of Student Responses to a Collaborative Problem Solving Task

We would like to invite your school to take part in a research study. In particular, Sofia Eleftheriadou, a research student from the University of Manchester, is conducting research to find out more about young people's collaborative problem solving skills and we will be very grateful if you can help. This document provides more information about the study and what this will entail for your school.

What information are we collecting, and why?

The study aims to find out how young people aged 14 to 16 years old respond to a collaborative problem solving task. We are asking students in Years 9 to 11 to provide their agreement to some statements about collaboration, complete an online task alone and as part of a group of two, and discuss with the researcher the responses they have given and the way they approached the task. This will allow us to understand how young people make decisions and answer a collaborative problem solving task. We are interested in particular in comparing students' answering the task alone and as part of a small group (dyad) because we believe that there might be a difference between the two.

What will the school be required to do?

We are asking that you send out a project information letter and a consent form to parents which we will send you. Students who are interested in participating in the research will need to have consent forms signed by their parents. We would ask you collect those forms and inform the researcher about the number of students interested. The researcher will then visit your school at an appropriate time previously agreed with you in order to administer the task and interview the students.

During the visit, the researcher will administer the task in a laptop and will use a camera (both provided by the university) to video tape the interview and the task completion. The administration of the task and the interview will need to take place in a quiet room at school during school time. The researcher will meet with participating students twice for the completion of the task and the interview and this will take place face-to-face, first on one-to-one basis and then in groups of two, depending on students' availability and school's programme. Each meeting is expected to take no more than 1 hour.

We are also asking you to provide some background information for each student participating (such as their eligibility for free schools meals, and their grades in science, maths and English) so that we can match their responses with other information about them (of course after securing parental and student consent).

How do we ensure this information is managed securely?

Student data will be treated with the strictest confidence. We will not be transferring any identifiable information outside the EU and will be taking appropriate measures to ensure it remains secure at all times. This will be achieved with the use of password protected transfer, and the use of secure servers when transferring data between the research student and supervisors. We will not use your students' or the name of any school in any report arising from the research. Student responses will be linked to the data provided by the school using a unique ID.

Data for this project is being used in line with the University of Manchester Research Ethics Committee and the GDPR regulations.

Will the outcomes of the research be published?

It is expected that the outcomes of this study will be part of the student thesis submitted for the PhD Education qualification. They could also eventually be incorporated into publications in peer reviewed scientific journals, conference proceedings and presentations.

We thank you in advance for your help.

Yours sincerely,

Sofia Eleftheriadou

Sofia.eleftheriadou@manchester.ac.uk

10.10 Appendix 10: Information sheets and consent forms for students and parents/guardians

MANCHESTER
1824
The University of Manchester

An Exploration of Student Responses to a Collaborative Problem Solving Task

Student consent form

If you are happy to participate please put a tick next to the statements below.

Activities	Tick
I have read the information for the study and had the opportunity to ask questions.	
I understand that I can say 'no' if I no longer wish to participate in the study.	
I agree to my task responses in the computer and discussions with the researcher being video recorded.	
I agree to the use of anonymous quotes.	
I agree that any data collected may be published in anonymous form in academic books or journals.	
I agree to take part in this study	

Data Protection: The personal information we collect and use to conduct this research will be processed in accordance with data protection law as explained in the Participant Information Sheet and the Privacy Notice for Research Participants.

________________________ ____________________ ______________
Your name Signature Date

_______________________ ______________________ _____________
Name of the researcher Signature Date

An exploration of student responses to a collaborative problem solving task

Who is Conducting the Research?

My name is Sofia and I work as a researcher at the University of Manchester. I would like to invite you to take part in our research study about students working together to solve a task.

Before you decide if you wish to take part, please make sure that you understand:
1. Why the research is being done
2. What your involvement in the project will be

Take your time to read through this information sheet before you decide if you wish to take part. Ask as many questions as you wish.

What is the Purpose of the Research?

The aim of the research is to explore how students respond to a problem solving task by collaborating with others. We think that there is a difference between completing a problem solving task alone and in pairs. The results of this study will help us understand the difference in students' thinking processes and responses when completing the task individually and when working in small groups.

Why Have I Been Asked to Take Part?

We have asked you to take part because you are aged 14-16 and attend school.

What Would I Be Asked to Do if I Take Part?

If you want to take part, we will ask that you volunteer to take part in the following:

- At the start of the interview, I will ask you to report how much you agree with some statements.

These will be simple sentences that I would like you to first read and report your agreement with and then I discuss with you how you understand them and why you answered the way you did. There will be no right or wrong answers.

- Complete an online task (alone)

In the online task you will join a team of students that take part in a school contest. The aim will be to discuss in the team about the best way to complete the task. You will be able to talk to other team members by selecting some messages through a chat space. After completing the task, I would like to go through your responses and discuss them with you.

- Complete an online task (in a group of two)

You and another student from your school will work together on the same online task. After completing the task, I will go through your responses and discuss them with you.

I will first meet you for an interview alone and then in a group of two (or the opposite). I hope to video record you using a camera when completing the task and the discussion. I also hope to record the computer screen when you complete the task. The video will help me learn more about how you respond to the task. Nobody will see the video recordings except for me and my supervisors. It is important to know that there are **no right or wrong answers for this task**.

I would also like to ask you about your age and language spoken at home and your school about your previous grades in maths, science and English, and your eligibility for free school meals. Whatever you say to me will be kept safe. Nobody is going to know about your name or your school and this study will not affect in any way your school grades or results.

How long is the Study?

I will meet with you twice and each meeting will take no more than 1 hour.

Where will the Study Take Place?

The study will take place in your school during school time.

Will my Participation in the Study be Confidential?

Only the researcher and supervisors will have access to data and we will ensure it is kept safe and secure in accordance with the General Data Protection Regulation (GDPR) and Data Protection Act 2018. If you would like to know more about how we keep your information safe and comply with the law, please read through our Privacy Notice for Research or discuss the privacy notice with your parent/guardian.

The University of Manchester will protect the information about you as we are called the Data Controller (this means we have to protect your information by law). To do this we have a number of safeguards in place such as policies and procedures. All researchers have received training to do this and we will make sure that they keep your information safe.

We will make sure that no one knows you have chosen to take part in the study and will also not share any information you have given to us. To do this we will use a process called anonymising, which means that we will generate a secret code for you and make sure that your name is stored in a different place to the rest of the information you give us. We will also keep the information you give us for less than 5 years and then it will be safely destroyed.

You have many rights under the new data protection laws and can request to see any of the information you have shared with us. This is called a Subject Access Request and if you would like to know more about your rights, please read through the Privacy Notice for Research or discuss it with your parent/guardian.

Do I Have to Take Part?

It is completely up to you if you wish to take part in the study. Make sure you think carefully and consider all the information contained in this sheet before you decide.

After you have decided you will be asked to sign an assent form that shows you understand and agree to take part in the research. Your parent/guardian will do the same (and sign a consent form) if they also agree for you to take part.

What if I Change my Mind?

You are free to withdraw from the study at any point without having to give a reason. You can ask for your data to be removed from analysis after it is collected by providing your unique id number that it will be given to you. Please remember that your data will be anonymised and you will not be identified in any way.

Who is Organising and Approving the Research?

The research is being sponsored by the University of Manchester and the Economic and Social Research Council.

The research has also been approved by the School of Environment, Education and Development Division/School Committee [Reference number: 2019-5841-10069], a group of people who work to protect your safety, rights, wellbeing and dignity.

What Do I Do Now?

If you have any questions relating to the information contained in this sheet, please let me know:

Researcher: Sofia Eleftheriadou, sofia.eleftheriadou@manchester.ac.uk

Research Supervisor: Maria Pampaka, maria.pampaka@manchester.ac.uk

Thank you for reading this!

MANCHESTER
1824
The University of Manchester

An Exploration of Student Responses to a Collaborative Problem Solving Task

Opt-in form

I hereby give my consent for Sofia Eleftheriadou to undertake this research by allowing my child to be interviewed and that any information collected about my child will be kept in confidence and for research purposes only. In particular I give consent for the following:

Activities	Please tick the box
I have read the information for the study and had the opportunity to consider the information provided.	
I understand that my child can say 'no' if he/she no longer wishes to participate in the study.	
I agree to my child's task responses in the computer and discussions with the researcher being video recorded.	
I agree to the use of anonymous quotes.	
I agree that any data collected may be published in anonymous form in academic books or journals.	
I agree for my child to take part in this study.	

Data Protection: The personal information we collect and use to conduct this research will be processed in accordance with data protection law as explained in the Participant Information Sheet and the [Privacy Notice for Research Participants].

____
Name of child

_______________________ _______________________ _______________________
Parent/gurdian's name Signature Date

_______________________ _______________________ _______________________
Name of the researcher Signature Date

Please return this signed form to the school.

MANCHESTER
1824
The University of Manchester

An Exploration of Student Responses to a Collaborative Problem Solving Task

Information about the project

Your child has been invited to take part in a research study run by the University of Manchester about collaborative problem solving skills of young people aged 14 to 16 years old. Before you decide whether you agree with your child to take part it is important for you to understand why the research is being done and what it will involve. Please read the following information about the project before you decide whether or not you agree for your child to take part.

Who will conduct the research?

Sofia Eleftheriadou – PhD student
C3.32, Ellen Wilkinson Building, Manchester Institute of Education, The University of Manchester.

What is the aim of the research?

The study aims to find out how young people aged 14 to 16 years old respond and make decisions in a collaborative problem solving task individually and as part of a small group.

Why have my child been chosen?

Your child's school is one of the educational establishments that have agreed to take part in this study.

What would my child be asked to do if he/she took part?

In these schools we are asking a few students in Years 9 to 11 to complete an online task individually and as part of a group of two and provide their agreement to some statements about collaboration. This will allow us to understand how young people respond and make decisions in completing a collaborative problem solving task as well as what are the differences between completing it alone or as part of a group. Completing the task will take about 30 minutes and this will be done during school time. After the completion of the task, the researcher would like to discuss with the students about their responses and their way of thinking in completing the task. These interviews will take place at school after the completion of the task and should not last more than 20 minutes.

Each session (of two) is not expected to take more than 1 hour and both will be conducted face-to-face with the researcher at school during school time. We prefer to video record the interviews and tasks that students complete for research purpose. We would also like to video record the computer screen when students complete the task.

If you agree for your child to take part in the research study, please sign the consent form that follows.

What happens to the data collected?

In order to undertake the research project we will need to collect some personal information/data about your child to link the responses from the task. In the interview, we will ask you child about his/her age and language spoken at home. If you agree for your child to take part, we will also ask school to provide us your child's previous grades in maths, science and English and his/her eligibility for free school meals. Any data collected will be fully anonymised (information will not enable identification of your child).

A video-recording of the interview and the task will be made so that it can then be transcribed and linked with the personal identifier provided by the school. The transcripts of the interview will be used to provide background information about the way students responded to the task. Excerpts from the anonymized transcripts may also be used in future research papers and presentations as part of the normal research dissemination process.

How is confidentiality maintained?

We will not use your child's name or the name of the school in any report arising from the research. All participants will remain anonymous and be given aliases when quoted in the thesis or any other writing. All video recordings and data will be stored in encrypted files and will only be accessed by the researcher and researcher's supervisors.

Only the researcher and supervisors will have access to this and we will ensure it is kept safe and secure in accordance with the General Data Protection Regulation (GDPR) and Data Protection Act 2018. If you would like to know more about how we keep your child's information safe and comply with the law, please read through our Privacy Notice for Research.

The University of Manchester will protect the information about your child as we are called the Data Controller (this means we have to protect your information by law). To do this we have a number of safeguards in place such as policies and procedures. All researchers have received training to do this and we will make sure that they keep your information safe.

We will make sure that no one knows your child has chosen to take part in the study and will also not share any information they have given to us. To do this we will use a process called anonymising, which means that we will generate a secret code for your child and make sure that his/her name is stored in a different place to the rest of the information he/she give us. We will also keep the information he/she give us for less than 5 years and then it will be safely destroyed.

You have many rights under the new data protection laws and can request to see any of the information you have shared with us. This is called a Subject Access Request and if you would like to know more about your rights, please read through the Privacy Notice for Research.

Will my child's data be used for future research?

No

What happens if I do not want my child to take part or if I change my mind?

It is up to you to decide whether or not your child can take part. If you do decide for your child to take part you will be given this information sheet to keep and be asked to sign a consent form (provided below). If you decide for your child to take part you are still free to withdraw your child's data at any time without giving a reason.

Will I be paid for participating in the research?

Participation is voluntary and there will be no financial payment.

What is the duration of the research?

There will be 2 face-to-face meetings of the researcher with your child. Each meeting is not expected to take more than 1 hour.

Where will the research be conducted?
The study will take place at the school during school time.

Disclosure and Barring Service (DBS) Check

The researcher who will interview your child have undergone a satisfactory DBS check.

Will the outcomes of the research be published?

The outcomes of the research will be published in a doctoral thesis and may also be used for publications in peer reviewed scientific journals, conference proceedings and presentations.

Contact for further information:
Researcher details: Sofia Eleftheriadou
Email: sofia.eleftheriadou@manchester.ac.uk

What if I want to make a complaint?

If you have a minor complaint then you need to contact the researcher named above or the supervisor Maria Pampaka, maria.pampaka@manchester.ac.uk, in the first instance.

If you wish to make a formal complaint or if you are not satisfied with the response you have gained from the researchers in the first instance then please contact: The Research Governance and Integrity Manager, Research Office, Christie Building, University of Manchester, Oxford Road, Manchester, M13 9PL, by emailing: research.complaints@manchester.ac.uk or by telephoning 0161 275 2674.

What Do I Do Now?

If you are happy for your child to take part in the research, **then please sign the consent form that follows and return it to the school.** Your child may withdraw at any time.

Thank you again for your support for this important study.

This Project Has Been Approved by the University of Manchester's Research Ethics Committee [Reference number: 2019-5841-10069]

10.11 Appendix 11: Cognitive interviewing – Interview protocol

Step 1: Opening the interview: *As a researcher, I am particularly interested in how 15-year-olds understand some questions and statements used in a student questionnaire. That is why I would like to go through each question together with you and hear more about how you understand and interpret them. I am not interested in right and wrong answers. The most important point here is to get an honest as possible answer about how you experienced the questions.*

Step 2: Completion of assessment

I would like to see how you understand an online task. Can we start with you answering the online task first? (Open task – start screen video) If you get stuck on anything, don't worry, move onto the next thing. At the end of the task, we'll discuss your responses.

(*I wait until the student completes the online task – 12 items*)

Step 3: Validation of the assessment

I will now go through each step of the task together and talk about how you understand it. I am not interested in right and wrong answers. The most important point here is to get an honest as possible answer about how you experienced the questions. We will watch a video that recorded the screen when you were completing the task, for every part of the task that you responded, I will pause the video and ask you some questions about how you understood it.

(I play the video, pause when student gives a response and ask the main probes.)

Main probes

1) How do you understand this part?
2) Can you explain why you have given this answer to this part?

Follow-ups

1) Can you tell me a little more about what the question means to you? Can you give me an example?
2) Can you tell me a little more about why you chose that answer? Can you describe a time when that happened?

Anticipated probes for each item

Organised around the four cognitive operations; comprehension, recall, judgement, response

Step 4: Open questions about the CPS assessment
What if the phrase chat box was an open chat box instead? Would it be different for you?
Which parts were hard/easy?
Which parts were like things you've done before in the classroom? Give an example.
Which parts were like things you've done before in your home or hobbies? Give an example.
Which parts are like things you have done before using technology, like gaming? Can you give me an example?
What would you change?

Step 5: Closing the interview: *This was the last question that I wanted to ask you, is there anything else that you would like to add? Do you have any other questions regarding our interview today that you would like to ask? Thank you very much for your participation.*

10.12 Appendix 12: List of initial codes

Table 10.9. Codes generated during initial coding phase

Initial codes (in alphabetical order)	**References**
addressing gaps	5
answering faster	6
approaching the task	2
asking for reasons	8
assigning roles	14
being confused	18
being kind	2
being nice - respectful- polite	5
being part of the situation	3
boosting confidence	2
building a common ground	8
changing strategy	14
changing the plan	6
checking everyone is on the same page	1
choosing option by accident	2
claiming remaining subject	1
communicating with team about approach	8
complaining	2
debating about response	1
deciding not to argue	3
deciding on a subject	4
deciding on one response	3
encouraging team	8
ensuring team is happy	5
evaluating alternative plan	3
evaluating progress	15
evaluating reasons	40
evaluating response options	34
feeling uncomfortable with a response	10
finding a fair way	2
finding a response obvious	13
finding a response rude	4
finding response options limited	1
finding the response immature	1
following plan	3
getting the task done quicker	19
getting the team back on track	1
giving a middle option	3
having more freedom	5

helping each other	21
identifying gaps	8
interpreting problem space	7
joking	2
keeping the group united	1
making a decision	4
misunderstanding instructions	1
not finding a response	4
not getting the choice to choose	4
not making sense	2
not realising violation in plan	10
noticing violation of plan	13
picking the best answer for the situation	3
praising a teammate	6
preferring a type-in response	6
prioritising team members	4
proposing plan	18
providing alternative response	23
providing clarification	12
providing reasons	4
recognising other person's opinion	8
recognising time limit in task	3
recognising time needs for finding a strategy	1
reflecting on a personal experience	2
reinforcing teamwork	1
satisficing	11
saving time	2
saying (or not) in real life	6
showing the violation in response option	1
staying on topic	3
taking feelings into account	2
taking leader's role	2
talking about what to do	10
thinking about finding a strategy	11
understanding different perspectives	5
understanding people's abilities	15
understanding teammates' strategy approaches	7
understanding the tone of responses	8
wasting time	8
wording response differently	9
working in or as a team	5

10.13 Appendix 13: Focused coding example

In the following example, I compared situations in which participants had responded correctly to item X6, with those who had not. Regardless of students' response in the previous item (X5), the second part of the task was programmed to start with a conflict between Alice and Zach. Both teammates said that they wanted to take the subject People. Students could then select one of the following chat messages as their response to Alice and Zach:

> -Nobody asked me what subject I want. Why should you guys choose first?
> -Can each of you explain why you want that subject?
> -Why are we wasting time arguing about this?
> -Alice and Zach, are you going to answer questions faster than you choose subjects?

The item's intent was to assess students' competence in discovering perspectives and abilities of team members, and the credited response was "Can each of you explain why you want that subject?" In addition to finding out teammates' reasons, two students found "Can each of you explain why you want that subject?" as the rational response to give, compared to the remaining options. Table 10.10 shows examples of the focused codes developed. For example, the initial codes 'finding a response pointless' and 'finding a response argumentative' were developed as a focused code 'evaluating responses.' Inspecting the responses of students who have not selected the credited response it was found that one of them was not feeling comfortable answering this item. Specifically, Anna felt that she was trying to lead her teammates and that made her feel uncomfortable. As an alternative response she would make a joke. Here, the implied responsibility to help teammates negotiate a solution when conflict arose, was what made Anna feel like she was leading her teammates, which eventually made her uncomfortable. In addition, a limitation of the task, made obvious by Anna's quote below, is that making a joke or answering in a humorous way was not something that has been offered by the assessment or accounted for by the framework. As shown in Table

10.10, the initial code ‘joking’ was developed as the focused code ‘understanding tone of the response.’

Table 10.10. Focused coding example

Focused codes	**Interview excerpts**
Evaluating responses	**<u>Excerpt 1:</u>** *I just thought, rather than arguing about who wants what, it would be better to **figure out which person is more suited** to which category. [...] Out of all those options the explaining why sounds like the **most appropriate one**. Because I found that the other ones, like, the top one [response] is more about my priorities, which I’d rather put what other people want to do first, and then I found the third one [response] **quite pointless** because it doesn’t really have **any relevance to anything,** and the last one [response], it’s quite sassy, it’s **not going to help the situation**. (Emily)* **<u>Excerpt 2:</u>** *I thought, well, at the end of the day, if I jumped and said the top one [response], that I want that subject, it would **create an argument and divide the group**. Then the third one [response], it doesn’t **resolve the issue** and the last one, I thought that wasn’t helping either. I thought the second one [response] was going to help the most. […] They both want to take People, it’s going to cause an argument over this one thing, so instead you need to **think reasonably** and think why do each one of them actually want this. So, I thought **the rational response** there. (Stephan)*
Feeling uncomfortable	*They are arguing because they both want to take one of the subjects. I mean**, it wasn’t super comfortable**, it just felt a bit, like **I am trying to lead them**. (Anna)*
Understanding tone of the response	*I would say ‘Take the other ones and I’ll take People’, but I probably **wouldn’t actually say that in a serious way**, like I would probably **be joking**. (Anna)*

10.14 Appendix 14: List of focused codes

Table 10.11. Codes generated during focused coding phase

Focused codes (in alphabetical order)	**Initial codes**
approaching the task	finding a fair way changing strategy asking for reasons assigning roles helping each other being part of the situation talking about what to do communicating about approach staying on topic
encouraging team	boosting confidence ensuring team is happy praising a teammate
evaluating progress	addressing gaps identifying gaps interpreting problem space not realising violation in plan noticing violation of plan
evaluating reasons	deciding on a subject making a decision
evaluating response options	finding the response immature finding a response obvious finding a response rude not making sense not finding a response preferring a type-in response finding response options limited having more freedom picking the best answer for the situation providing alternative response wording response differently showing the violation in response option
feeling uncomfortable	Complaining being confused not getting the choice to choose
getting the task done quicker	answering faster misunderstanding instructions recognising time needs for finding a strategy saving time wasting time recognising time limit in task

proposing plan	changing the plan evaluating alternative plan following plan getting the team back on track
providing clarification	building a common ground checking everyone is on the same page claiming remaining subject
satisficing	being nice - respectful- polite choosing option by accident
understanding teammates	understanding teammates' abilities understanding different perspectives understanding teammates' strategy approaches recognising other person's opinion
understanding the tone of responses	taking leader's role joking saying (or not) in real life reflecting on a personal experience
working in or as a team	being kind debating about response deciding not to argue giving a middle option keeping the group united prioritising team members reinforcing teamwork taking feelings into account

10.15 Appendix 15: Coding of articles in the CPS concepts review (Chapter 4/Paper 1)

Table 10.12. Coding of articles in the category CPS competence

Author (1st only)	Journal	Year	Country affiliation of 1st author	Type of research	Type of evidence	Educational setting	Content area	Sample size	Group mode	Method of data collection	Unit of analysis	Method of data analysis
Andrews-Todd	Computers in Human Behavior	2020	USA	empirical	quant	university	STEM	129	H-H	computer-simulated task	individual chats/actions	quantitative content analysis
Andrews-Todd	ETS Research Report Series	2019	USA	empirical	quant	secondary	STEM	10	H-H	computer-simulated task	individual chats/actions	quantitative content analysis
Andrews-Todd	International Journal of Testing	2019	USA	theoretical	NA	NA	NA	NA	NA	NA	NA	NA
Camacho-Morles	Learning and Individual Differences	2019	Australia	empirical	quant	secondary	mixed	200	H-H	computer-simulated task	individual chats/actions	Item Response Theory, Structural Equation Modeling
Care	Applied Measurement In Education	2016	Australia	theoretical	NA	NA	NA	NA	NA	NA	NA	NA
Care	Research And Practice In Technology Enhanced Learning	2014	Australia	empirical	quant	secondary	generic	4056	H-H	computer-simulated task	individual chats/actions	Item Response Theory
Cukurova	Computers & Education	2018	UK	empirical	quant	secondary, university	STEM	45	H-H	face-to-face task	individual and sequences of	Multi-modal learning

											non-verbal behaviour	analytics system
De Boeck	Frontiers in Psychology	2019	USA	empirical	quant	secondary	generic	986	H-A	computer-simulated task	individual chats/actions	Confirmatory Factor Analysis
Graesser	Psychological Science In The Public Interest	2018	USA	theoretical	NA	NA	NA	NA	NA	NA	NA	NA
Häkkinen	Teachers And Teaching: Theory And Practice	2017	Finland	theoretical	NA	NA	NA	NA	NA	NA	NA	NA
Harding	Journal Of Applied Measurement	2016	Australia	empirical	quant	secondary	mixed	3402	H-H	computer-simulated task	individual chats/actions	Item Response Theory
Harding	Aera Open	2017	Australia	empirical	quant	secondary	mixed	3004	H-H	computer-simulated task	individual chats/actions	Item Response Theory
Herborn	Computers In Human Behavior	2020	Luxembourg	empirical	quant	secondary	generic	748	H-A	computer-simulated task	individual chats/actions	Structural Equation Modeling
Herborn	Journal Of Educational Measurement	2017	Luxembourg	empirical	quant	secondary	generic	481	H-A	computer-simulated task	individual chats/actions	Cluster analysis
Herro	International Journal Of Stem Education	2017	USA	theoretical	NA	NA	NA	NA	NA	NA	NA	NA
Hsieh	Computers In Human Behavior	2002	USA	empirical	quant	secondary	STEM	120	H-H	computer-simulated task	individual chats/actions; group outcome	quantitative content analysis

Krkovic	European Journal Of Psychological Assessment	2016	Germany	empirical	quant	secondary	generic	483	H-A	computer-simulated task	individual chats/actions	quantitative content analysis
Li	Eurasia Journal Of Mathematics Science And Technology Education	2017	Taiwan	empirical	quant	secondary	STEM	52110	H-A	computer-simulated task	individual chats/actions	quantitative content analysis
K.-Y. Lin	Journal Of Computers In Education	2015	Taiwan	empirical	quant	secondary	STEM	222	H-A	computer-simulated task	individual chats/actions	quantitative content analysis
Nouri	International Journal Of Emerging Technologies In Learning	2017	Sweden	empirical	quant	secondary	generic	24	H-H	computer-simulated task	individual chats/actions	quantitative content analysis
Oliveri	International Journal of Testing	2019	USA	theoretical	NA	NA	NA	NA	NA	NA	NA	NA
Oliveri	ETS Research Report Series	2017	USA	theoretical	NA	NA	NA	NA	NA	NA	NA	NA
O'Neil	Assessment In Education: Principles, Policy & Practice	2003	USA	theoretical	NA	NA	NA	NA	NA	NA	NA	NA
Polyak	Frontiers in Psychology	2017	USA	empirical	quant	secondary	generic	159	H-A	computer-simulated task	individual chats/actions	Bayesian Evidence Tracing
Rosen	Technology, Knowledge And Learning	2014	USA	empirical	quant	secondary	generic	179	H-H; H-A	computer-simulated task	individual chats/actions	quantitative content analysis

Rosen	International Journal Of Artificial Intelligence In Education	2015	USA	empirical	quant	secondary	generic	179	H-H; H-A	computer-simulated task	individual chats/actions	quantitative content analysis
Rosen	Research And Practice In Technology Enhanced Learning	2014	USA	empirical	quant	secondary	generic	179	H-H; H-A	computer-simulated task	individual chats/actions	quantitative content analysis
Scoular	Educational Assessment	2019	Australia	empirical	quant	secondary	mixed	1210	H-H	computer-simulated task	individual chats/actions	Item Response Theory
Scoular	Computers in Human Behavior	2020	Australia	empirical	quant	secondary	generic	3010	H-H	computer-simulated task	individual chats/actions	Item Response Theory
Scoular	Journal Of Educational Measurement	2017	Australia	theoretical	NA	NA	NA	NA	NA	NA	NA	NA
Siddiq	Computers In Human Behavior	2017	Norway	empirical	quant	secondary	language	11	H-H	computer-simulated task	individual chats/actions	sequential analysis
Stadler	Journal of Intelligence	2019	Germany	empirical	quant	secondary	generic	748	H-A	computer-simulated task	individual chats/actions	Structural Equation Modeling
Sun	Computers & Education	2020	USA	empirical	quant	secondary	STEM	33	H-H	computer-simulated task	individual chats/actions	Principal Component Analysis
Vista	Computers In Human Behavior	2017	Australia	empirical	quant	secondary	generic	1214	H-H	computer-simulated task	sequences of group chats/actions	network analysis

Webb	Education And Information Technologies	2015	UK	theoretical	NA	NA	NA	NA	NA	NA	NA	NA
Yuan	Frontiers in Psychology	2019	China	empirical	quant	secondary	generic	434	H-H	computer-simulated task	Individual and sequences of chats/actions	Item Response Theory

Notes: In the category ‘Group mode’, the following abbreviations have been used: H-H=human-to-human approach, H-A=human-to-computer-simulated agent/partner approach.

Table 10.13. Coding of articles in the category CPS practice

Author (1st only)	Journal	Year	Country affiliation of 1st author	Type of research	Type of evidence	Educational setting	Content area	Sample size	Group mode	Method of data collection	Unit of analysis	Method of data analysis
Albert	Journal Of Mathematics Education At Teachers College	2013	USA	theoretical	NA	NA	NA	NA	NA	NA	NA	NA
Cáceres	Interactive Learning Environments	2018	Chile	empirical	quant	primary	STEM	75	H-H	pre- and post-test	individual scores	t-test
Cai	Educational Technology Research And Development	2016	China	empirical	quant	primary	STEM	21	H-H	computer-simulated task, pre- and post-test, questionnaire	individual scores, events	t-test, content analysis
Chao	Eurasia Journal Of Mathematics Science And Technology Education	2018	Taiwan	empirical	quant	primary	STEM	16	H-H	pre- and post-test	individual scores	t-test
Chen	Technology, Knowledge and Learning	2020	Japan	empirical	quant	secondary	STEM	31	H-H	pre-and post-test, questionnaire	individual scores	t-test
Cho	British Journal Of Educational Technology	2017	South Korea	empirical	quant	secondary	geography	101	H-H	pre- and post-questionnaire	individual scores	t-test
Cukurova	Oxford Review Of Education	2018	UK	theoretical	NA	NA	NA	NA	NA	NA	NA	NA

Gu	Journal Of Computer Assisted Learning	2019	China	empirical	quant	university	generic	49	H-H	computer-simulated task, pre-and post-test, questionnaire	individual scores, single utterances	t-test, quantitative content analysis
Huang	Computers & Education	2017	Taiwan	empirical	quant	primary	STEM	64	H-H	pre-and post-test, questionnaire	individual scores	t-test
P.-C. Lin	Interactive Learning Environments	2020	Taiwan	empirical	quant	university	STEM	84	H-H	computer-simulated task, pre- and post-test	individual scores, single utterances	t-test, quantitative content analysis, sequential analysis
Merrill	Educational Technology Research And Development	2002	USA	theoretical	NA	NA	NA	NA	NA	NA	NA	NA
Rosen	Computers in Human Behavior	2020	USA	empirical	quant	secondary	STEM	180	H-H	pre- and post-test	individual scores	t-test
Rosen	Journal Of Educational Measurement	2017	USA	empirical	quant	secondary	STEM	220	H-H; H-A	pre- and post-test	individual scores	t-test
Slof	Instructional Science: An International Journal Of The Learning Sciences	2012	The Netherlands	empirical	quant	secondary	economics	102	H-H	computer-simulated task	individual scores, episode	t-test, quantitative content analysis

Slof	Computers In Human Behavior	2012	The Netherlands	empirical	quant	secondary	economics	102	H-H	computer-simulated task	individual scores, single utterance	t-test, quantitative content analysis
Tawfik	Technology, Knowledge And Learning	2014	USA	empirical	quant	university	generic	22	H-H	computer-simulated task	single utterance	quantitative content analysis
Wu	Journal of Educational Computing Research	2020	Taiwan	empirical	quant	secondary	generic	68	H-H	computer-simulated task	single utterance	quantitative content analysis, sequential analysis

Notes: In the category ‘Group mode’, the following abbreviations have been used: H-H=human-to-human approach, H-A=human-to-computer-simulated agent/partner approach.

Table 10.14. Coding of articles in the category CPS interaction

Author (1st only)	Journal	Year	Country affiliation of 1st author	Type of research	Type of evidence	Educational setting	Content area	Sample size	Group mode	Method of data collection	Unit of analysis	Method of data analysis
Chan	Zdm - Mathematics Education	2017	Australia	empirical	qual	secondary	STEM	50	H-H	face-to-face task	event	content analysis
Chang	Computers & Education	2017	Taiwan	empirical	quant	secondary	STEM	30	H-H	computer-simulated task	sequence of utterances	sequential analysis
Chang	Journal Of Computer Assisted Learning	2017	Taiwan	empirical	quant	secondary	STEM	83	H-H	computer-simulated task	sequence of utterances	sequential analysis
Ding	Computers & Education	2009	The Netherlands	empirical	quant	secondary	STEM	6	H-H	computer-simulated task	single utterance	sequential analysis
Karabulut-Ilgu	Journal Of Educational Computing Research	2018	USA	empirical	mixed	university	STEM	64	H-H	face-to-face task, interview	single utterance, episode	quantitative content analysis, qualitative description
Kumpulainen	Journal Of Experimental Education	2003	Finland	empirical	qual	primary	STEM	20	H-H	face-to-face task	single utterance, episode	qualitative description

Notes: In the category 'Group mode', the following abbreviation has been used: H-H=human-to-human approach.

10.16 Appendix 16: Existing frameworks of CPS competence

Table 10.15. Existing frameworks of collaborative problem-solving competence

Framework	Main intended outcome(s)	Components of CPS competence
Framework for teachable CPS skills (Hesse et al., 2015)	Definition of 21st century skills Development of assessment and teaching approaches	Social: participation, perspective taking, social regulation Cognitive: task regulation, knowledge building
Programme for the International Student Assessment (PISA) 2015 CPS framework (OECD, 2017a)	Large-scale educational assessment of students' cognitive and non-cognitive skills Help governments shape their education policy	Collaboration processes: Establishing and maintaining shared understanding, taking appropriate action to solve the problem, establishing and maintaining team organisation Individual problem-solving processes: Exploring and understanding, representing and formulating, planning and executing, monitoring and reflecting
In-task assessment framework CPS Ontology (Andrews-Todd & Kerr, 2019)	Provide a theory-driven representation of the skills associated with CPS Provide guiding principles for CPS assessment and support in assessing proficiency in complex skills derived from high-granularity log data	Social: maintaining communication, sharing information, establishing shared understanding, negotiating Cognitive: exploring and understanding, representing and formulating, planning, executing, monitoring
Australian Council for Educational Research (ACER) Framework for Collaboration (Scoular, Duckworth, et al., 2020)	Establish a common terminology for describing collaboration in the context of problem solving Providing a structure for the assessment and teaching of collaboration	Building shared understanding: communicate with others, pool resources and information, negotiate roles and responsibilities Collectively contributing: participate in the group, recognise contributions of others, engage with role and responsibilities Regulating: ensure own contributions are constructive, resolve differences, maintain shared understanding, adapt

		behaviour and contributions for others
Generalised CPS competency model (Sun et al., 2020)	Provide a model to be used in both remote and face-to-face CPS settings to assess human-human interactions	Constructing shared knowledge: sharing understanding, establishing common ground Negotiation and coordination: responding to others' questions/ideas, monitoring execution Maintaining team function: fulfilling individual roles on the team, taking initiatives to advance collaboration processes
Holistic framework for enhancing education and workplace success (Camara et al., 2015)	Definition of CPS as composite construct, composed of skills	Problem solving, Communication, Behaviour feature identification, maintaining shared understanding, engagement, strategy, evaluate

10.17 Appendix 17: Item fit statistics for CPS competence overall measure

Table 10.16. Item measures and fit statistics for the overall CPS competence scale with 117 items (highlighted with grey are the Xandar items)

Item	Total score	Total count	Measure	Model S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	Item	Total score	Total count	Measure	Model S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
1	746	924	-1.09	0.09	0.99	-0.30	0.97	-0.40	60	764	906	-1.35	0.10	0.92	-1.30	0.89	-1.20
2	667	924	-0.55	0.08	0.93	-2.00	0.86	-2.70	61	526	906	0.19	0.07	0.90	-4.00	0.89	-3.40
3	629	924	-0.32	0.08	0.92	-2.50	0.88	-2.80	62	582	906	-0.11	0.07	0.83	-6.00	0.77	-6.20
4	348	924	1.12	0.07	0.95	-2.10	0.96	-1.00	63	576	906	-0.08	0.07	1.02	0.60	0.99	-0.20
5	770	924	-1.28	0.09	1.05	0.90	1.22	2.50	64	630	906	-0.38	0.08	0.86	-4.20	0.82	-4.00
6	697	924	-0.74	0.08	1.08	1.80	1.14	2.30	65	780	906	-1.50	0.10	1.06	0.90	1.15	1.40
7	549	923	0.10	0.07	0.97	-1.00	0.97	-0.90	66	723	906	-1.01	0.09	0.96	-0.80	0.88	-1.70
8	469	923	0.50	0.07	1.01	0.60	1.02	0.70	67	656	906	-0.54	0.08	0.99	-0.20	0.97	-0.50
9	732	923	-0.99	0.09	0.88	-2.60	0.78	-3.40	68	698	906	-0.82	0.08	0.93	-1.60	0.93	-1.00
10	649	923	-0.45	0.08	0.94	-1.60	0.91	-1.90	69	644	888	-0.55	0.08	0.82	-5.10	0.72	-5.60
11	666	923	-0.55	0.08	1.11	3.00	1.18	3.30	70	364	888	0.99	0.07	1.17	6.30	1.23	6.20
12	798	923	-1.55	0.10	0.99	-0.10	1.05	0.60	71	506	888	0.25	0.07	1.01	0.50	1.02	0.50
13	2087	923	-0.60	0.04	1.02	0.50	1.10	1.70	72	711	888	-1.02	0.09	0.95	-1.00	0.88	-1.60
14	655	923	-0.48	0.08	1.05	1.30	1.08	1.60	73	671	888	-0.73	0.08	0.84	-4.00	0.76	-4.10
15	892	923	-3.14	0.19	0.95	-0.30	0.56	-2.20	74	569	888	-0.10	0.08	0.93	-2.50	0.91	-2.40
16	545	916	0.11	0.07	0.93	-2.90	0.91	-2.90	75	685	888	-0.83	0.08	1.04	0.80	1.08	1.30
17	284	916	1.45	0.08	1.08	2.60	1.19	3.80	76	454	888	0.52	0.07	1.00	-0.20	1.00	-0.10
18	577	916	-0.06	0.07	0.93	-2.40	0.92	-2.10	77	717	888	-1.07	0.09	0.90	-2.00	0.82	-2.50
19	610	916	-0.24	0.07	1.01	0.30	1.03	0.70	78	740	884	-1.37	0.10	1.01	0.20	1.05	0.60
20	580	916	-0.08	0.07	0.95	-1.60	0.95	-1.40	79	625	884	-0.53	0.08	0.92	-2.20	0.89	-2.30
21	322	916	1.24	0.07	1.09	3.20	1.18	4.30	80	592	884	-0.33	0.08	0.99	-0.30	0.98	-0.40
22	1094	916	0.76	0.04	1.13	3.10	1.25	4.30	81	533	884	-0.01	0.07	0.95	-1.80	0.94	-1.80

23	559	905	0.00	0.07	1.01	0.40	1.03	0.90	82	656	884	-0.73	0.08	0.97	-0.70	0.95	-0.90
24	244	905	1.67	0.08	1.08	2.20	1.20	3.50	83	474	884	0.31	0.07	0.96	-1.70	0.95	-1.70
25	443	905	0.60	0.07	1.13	5.60	1.16	5.40	84	761	884	-1.58	0.10	0.88	-1.80	0.69	-3.50
26	428	905	0.67	0.07	0.94	-2.60	0.94	-2.00	85	488	884	0.23	0.07	1.19	7.30	1.25	7.70
27	583	905	-0.13	0.07	0.86	-4.80	0.82	-5.00	86	735	884	-1.33	0.09	0.97	-0.50	0.95	-0.60
28	688	905	-0.77	0.08	0.93	-1.60	0.91	-1.60	87	155	884	2.20	0.09	1.05	1.00	1.21	2.40
29	499	905	0.31	0.07	0.93	-3.10	0.92	-2.70	88	397	883	0.71	0.07	1.08	3.20	1.09	2.90
30	818	905	-1.96	0.12	0.87	-1.50	0.61	-3.60	89	425	883	0.56	0.07	1.12	4.90	1.14	4.60
31	1251	905	-0.22	0.05	0.96	-0.80	0.96	-0.80	90	1186	879	-0.11	0.05	0.98	-0.60	0.93	-1.00
32	517	882	0.15	0.07	0.87	-5.40	0.85	-4.90	91	452	879	0.41	0.07	1.12	4.90	1.14	4.70
33	544	882	0.00	0.07	1.01	0.50	1.01	0.40	92	1202	879	-0.19	0.05	1.11	2.50	1.14	2.10
34	411	882	0.70	0.07	0.96	-1.90	0.95	-1.60	93	988	879	0.21	0.05	0.87	-3.60	0.85	-3.80
35	365	882	0.94	0.07	1.04	1.70	1.07	2.00	94	189	879	2.05	0.06	1.13	1.60	2.35	5.80
36	594	882	-0.28	0.08	1.05	1.50	1.07	1.50	95	1158	877	-0.05	0.05	1.01	0.30	0.97	-0.40
37	416	882	0.67	0.07	1.13	5.30	1.13	4.10	96	1214	877	0.68	0.04	1.07	1.70	1.11	2.20
38	378	882	1.55	0.05	1.06	1.20	1.10	0.90	97	1108	877	-0.39	0.06	0.94	-1.40	0.95	-1.20
39	517	925	0.29	0.07	1.05	2.20	1.05	1.70	98	699	877	-1.07	0.09	0.94	-1.20	0.94	-0.80
40	1140	925	0.06	0.05	0.95	-1.30	0.92	-1.70	99	411	877	1.57	0.05	1.00	0.10	1.16	2.20
41	196	919	2.05	0.08	1.01	0.20	1.16	2.10	100	2072	877	-0.48	0.04	1.20	3.30	1.23	1.40
42	293	919	1.44	0.08	1.11	3.40	1.22	4.40	101	472	877	0.30	0.07	0.87	-5.70	0.86	-5.20
43	478	919	0.47	0.07	1.09	3.80	1.11	3.60	102	630	877	1.33	0.06	1.18	4.00	1.24	5.40
44	595	919	-0.14	0.07	0.97	-0.90	0.95	-1.20	103	455	877	0.39	0.07	0.99	-0.30	0.99	-0.40
45	341	919	1.17	0.07	0.99	-0.30	1.01	0.20	104	525	867	-0.01	0.07	0.91	-3.50	0.88	-3.50
46	563	919	0.03	0.07	0.91	-3.30	0.88	-3.40	105	656	867	-0.81	0.08	0.96	-0.90	0.91	-1.50
47	413	919	0.80	0.07	0.80	-9.00	0.77	-7.70	106	125	867	2.46	0.10	1.08	1.30	1.46	4.20
48	1718	919	0.20	0.03	1.33	6.40	1.46	5.20	107	482	867	0.22	0.07	0.95	-2.10	0.94	-1.80
49	243	919	1.73	0.08	1.18	4.60	1.44	6.80	108	628	867	-0.62	0.08	0.85	-4.20	0.76	-4.90

50	545	919	0.13	0.07	0.87	-5.20	0.83	-5.30
51	830	917	-1.98	0.12	0.90	-1.20	0.71	-2.50
52	441	917	0.65	0.07	0.97	-1.10	0.96	-1.30
53	650	917	-0.46	0.08	0.94	-1.70	0.95	-0.90
54	680	917	-0.65	0.08	0.83	-4.50	0.74	-5.00
55	682	917	-0.66	0.08	0.94	-1.60	0.93	-1.20
56	714	917	-0.88	0.08	1.17	3.50	1.40	5.20
57	694	917	-0.74	0.08	0.81	-5.00	0.68	-5.80
58	2681	917	-0.45	0.04	1.13	2.80	1.18	3.00
59	609	906	-0.26	0.08	1.07	2.30	1.10	2.10
109	827	867	0.56	0.04	1.01	0.30	0.98	-0.30
110	1140	867	0.78	0.03	1.10	2.30	1.34	4.40
111	321	846	1.04	0.08	1.13	4.70	1.17	4.40
112	518	846	-0.04	0.08	1.02	0.60	1.04	1.10
113	172	846	2.01	0.09	1.07	1.40	1.33	4.10
114	612	846	-0.62	0.08	0.83	-4.70	0.74	-5.40
115	729	846	-1.58	0.10	1.02	0.30	1.03	0.40
116	253	835	1.42	0.08	1.16	4.50	1.30	5.50
117	734	835	0.69	0.05	1.15	3.90	1.20	3.60

Person separation: 2.70 Person reliability: 0.88

Item separation: 12.26 Item reliability: 0.99

10.18 Appendix 18: Item fit statistics for three sub-scales

Table 10.17. Item measures and fit statistics for the sub-scale “Shared understanding” with 61 items (highlighted with grey are the Xandar items)

Item order	Item	Total score	Total count	Measure	Model S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
1	2	667	924	-0.54	0.08	0.96	-1.20	0.92	-1.50
2	4	348	924	1.18	0.07	0.97	-1.00	0.98	-0.60
3	5	770	924	-1.29	0.09	1.08	1.40	1.33	3.30
4	6	697	924	-0.73	0.08	1.11	2.50	1.20	2.90
5	8	469	923	0.55	0.07	1.02	0.80	1.02	0.70
6	9	732	923	-0.99	0.09	0.88	-2.50	0.76	-3.30
7	12	798	923	-1.56	0.10	0.99	-0.10	1.01	0.20
8	13	2087	923	-0.63	0.04	1.04	0.80	1.07	1.10
9	16	545	916	0.14	0.07	0.95	-2.00	0.93	-1.90
10	17	284	916	1.53	0.08	1.13	3.80	1.27	4.90
11	18	577	916	-0.04	0.07	0.94	-2.00	0.94	-1.40
12	21	322	916	1.31	0.07	1.11	3.70	1.22	4.60
13	23	559	905	0.02	0.07	1.05	1.70	1.07	1.60
14	32	517	882	0.17	0.07	0.88	-4.60	0.84	-4.60
15	33	544	882	0.02	0.08	1.02	0.80	1.03	0.80
16	34	411	882	0.75	0.07	0.97	-1.00	0.97	-0.90
17	35	365	882	1.00	0.07	1.08	2.90	1.10	2.60
18	36	594	882	-0.27	0.08	1.06	1.70	1.06	1.20
19	40	1140	925	0.07	0.05	1.00	-0.10	0.96	-0.80
20	41	196	919	2.15	0.09	1.02	0.50	1.23	2.70
21	46	563	919	0.05	0.07	0.92	-2.70	0.88	-2.90
22	51	830	917	-2.02	0.12	0.90	-1.20	0.70	-2.20
23	52	441	917	0.70	0.07	0.98	-0.70	0.97	-0.80
24	53	650	917	-0.46	0.08	0.97	-0.90	1.03	0.50
25	54	680	917	-0.66	0.08	0.82	-4.70	0.72	-4.70
26	55	682	917	-0.67	0.08	0.95	-1.10	0.99	-0.10
27	56	714	917	-0.89	0.09	1.21	4.30	1.47	5.50
28	57	694	917	-0.75	0.08	0.81	-4.80	0.68	-5.20
29	58	2681	917	-0.50	0.04	1.20	4.00	1.34	4.90
30	61	526	906	0.22	0.07	0.90	-3.80	0.89	-3.10
31	62	582	906	-0.09	0.08	0.82	-6.10	0.74	-6.20
32	63	576	906	-0.06	0.08	1.04	1.20	1.01	0.30
33	64	630	906	-0.38	0.08	0.86	-4.10	0.83	-3.30
34	65	780	906	-1.53	0.10	1.08	1.30	1.22	1.90
35	69	644	888	-0.55	0.08	0.81	-5.30	0.69	-5.70
36	70	364	888	1.05	0.07	1.18	6.20	1.27	6.10
37	71	506	888	0.27	0.07	1.04	1.30	1.05	1.30
38	72	711	888	-1.04	0.09	0.97	-0.60	0.91	-1.10

39	73	671	888	-0.73	0.08	0.83	-4.20	0.74	-4.00
40	74	569	888	-0.08	0.08	0.94	-1.90	0.92	-1.80
41	75	685	888	-0.84	0.09	1.06	1.40	1.22	2.80
42	79	625	884	-0.53	0.08	0.94	-1.80	0.93	-1.30
43	80	592	884	-0.32	0.08	1.00	0.10	0.99	-0.20
44	81	533	884	0.02	0.07	0.97	-1.10	0.95	-1.20
45	83	474	884	0.34	0.07	0.96	-1.60	0.94	-1.70
46	87	155	884	2.31	0.09	1.09	1.50	1.32	3.10
47	90	1186	879	-0.11	0.05	1.00	0.10	0.94	-0.60
48	91	452	879	0.45	0.07	1.14	5.30	1.18	4.90
49	92	1202	879	-0.20	0.05	1.18	3.90	1.21	2.70
50	95	1158	877	-0.05	0.05	1.05	1.10	1.01	0.20
51	97	1108	877	-0.39	0.06	0.97	-0.70	0.97	-0.70
52	104	525	867	0.01	0.08	0.91	-3.20	0.87	-3.50
53	105	656	867	-0.82	0.09	0.98	-0.40	0.92	-1.10
54	106	125	867	2.57	0.10	1.11	1.70	1.68	5.20
55	107	482	867	0.25	0.07	0.97	-1.20	0.98	-0.60
56	108	628	867	-0.62	0.08	0.84	-4.30	0.74	-4.60
57	109	827	867	0.61	0.04	1.06	1.50	1.07	1.10
58	111	321	846	1.11	0.08	1.14	4.60	1.19	4.20
59	112	518	846	-0.02	0.08	1.05	1.50	1.09	2.00
60	113	172	846	2.11	0.09	1.11	2.20	1.47	5.00
61	114	612	846	-0.61	0.08	0.83	-4.60	0.75	-4.50

Person separation: 1.98 Person reliability: 0.80

Item separation: 11.63 Item reliability: 0.99

Table 10.18. Item measures and fit statistics for the sub-scale “Taking actions” with 26 items (highlighted with grey are the Xandar items)

Item order	Item	Total score	Total count	Measure	Model S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
1	1	746	924	-1.12	0.09	0.98	-0.50	0.99	0.00
2	3	629	924	-0.32	0.08	0.92	-2.40	0.89	-2.10
3	11	666	923	-0.56	0.08	1.12	3.10	1.18	2.80
4	14	655	923	-0.49	0.08	1.07	1.90	1.13	2.20
5	15	892	923	-3.27	0.19	0.92	-0.40	0.51	-2.10
6	20	580	916	-0.07	0.08	0.95	-1.50	0.96	-0.90
7	26	428	905	0.72	0.07	0.95	-1.80	0.99	-0.10
8	27	583	905	-0.12	0.08	0.88	-3.90	0.82	-4.20
9	31	1251	905	-0.24	0.05	0.89	-2.60	0.87	-2.10
10	38	378	882	1.69	0.05	0.98	-0.30	1.04	0.30
11	42	293	919	1.52	0.08	1.09	2.80	1.20	3.60
12	44	595	919	-0.12	0.08	0.98	-0.70	0.94	-1.30
13	48	1718	919	0.20	0.04	0.98	-0.30	0.94	-0.60
14	49	243	919	1.83	0.08	1.20	4.90	1.53	7.00
15	50	545	919	0.16	0.07	0.86	-5.40	0.82	-4.60
16	66	723	906	-1.01	0.09	0.97	-0.60	0.94	-0.60
17	67	656	906	-0.53	0.08	0.99	-0.10	1.00	0.10
18	68	698	906	-0.82	0.09	0.95	-1.10	0.90	-1.30
19	76	454	888	0.57	0.07	1.00	0.00	0.99	-0.30
20	77	717	888	-1.08	0.09	0.91	-1.80	0.79	-2.60
21	88	397	883	0.76	0.07	1.07	2.60	1.11	2.70
22	93	988	879	0.24	0.05	0.91	-2.50	0.87	-2.90
23	94	189	879	2.24	0.07	1.04	0.60	2.29	4.80
24	98	699	877	-1.10	0.09	0.91	-1.80	0.89	-1.30
25	100	2072	877	-0.50	0.04	1.06	1.00	1.18	0.90
26	102	630	877	1.43	0.06	1.23	5.00	1.28	5.80
		Person separation: 1.27				Person reliability: 0.62			
		Item separation: 13.94				Item reliability: 0.99			

Table 10.19. Item measures and fit statistics for the sub-scale “Team organisation” with 30 items (highlighted with grey are the Xandar items)

Item order	Item	Total score	Total count	Measure	Model S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
1	7	549	923	0.03	0.07	0.96	-1.40	0.98	-0.60
2	10	649	923	-0.53	0.08	0.95	-1.40	0.93	-1.30
3	19	610	916	-0.33	0.08	0.98	-0.60	1.03	0.70
4	22	1094	916	0.71	0.04	0.93	-1.60	0.99	-0.20
5	24	244	905	1.64	0.08	1.07	1.80	1.21	3.00
6	25	443	905	0.54	0.07	1.10	4.00	1.15	3.90
7	28	688	905	-0.87	0.08	0.95	-1.20	0.91	-1.30
8	29	499	905	0.24	0.07	0.93	-2.90	0.93	-2.10
9	30	818	905	-2.11	0.12	0.84	-1.90	0.59	-3.40
10	37	416	882	0.62	0.07	1.11	4.60	1.12	3.10
11	39	517	925	0.24	0.07	1.04	1.60	1.04	1.20
12	43	478	919	0.43	0.07	1.07	2.50	1.06	1.80
13	45	341	919	1.16	0.07	0.97	-1.10	0.98	-0.40
14	47	413	919	0.77	0.07	0.85	-6.00	0.82	-4.90
15	59	609	906	-0.33	0.08	1.04	1.10	1.04	0.80
16	60	764	906	-1.47	0.10	0.92	-1.30	0.92	-0.80
17	78	740	884	-1.46	0.10	1.01	0.10	1.03	0.40
18	82	656	884	-0.80	0.08	0.99	-0.10	0.97	-0.50
19	84	761	884	-1.67	0.10	0.88	-1.80	0.72	-3.00
20	85	488	884	0.18	0.07	1.13	5.10	1.19	5.30
21	86	735	884	-1.41	0.10	0.99	-0.20	0.93	-0.80
22	89	425	883	0.51	0.07	1.09	3.50	1.12	3.50
23	96	1214	877	0.64	0.04	0.95	-1.30	0.94	-1.30
24	99	411	877	1.55	0.05	0.99	-0.30	1.08	1.00
25	101	472	877	0.25	0.07	0.88	-4.90	0.88	-3.90
26	103	455	877	0.34	0.07	0.99	-0.50	1.00	0.10
27	110	1140	867	0.74	0.04	0.99	-0.10	1.01	0.20
28	115	729	846	-1.66	0.10	1.00	0.10	0.97	-0.30
29	116	253	835	1.38	0.08	1.16	4.40	1.27	4.60
30	117	734	835	0.65	0.05	1.09	2.20	1.12	2.20
		Person separation: 1.19				Person reliability: 0.58			
		Item separation: 12.76				Item reliability: 0.99			

10.19 Appendix 19: DIF analysis results for three sub-scales

This section presents results from the differential item functioning analysis conducted by gender for the three sub-scales of CPS competence. The analytical sample was n = 1,584 students (52.34% male and 47.66% female).

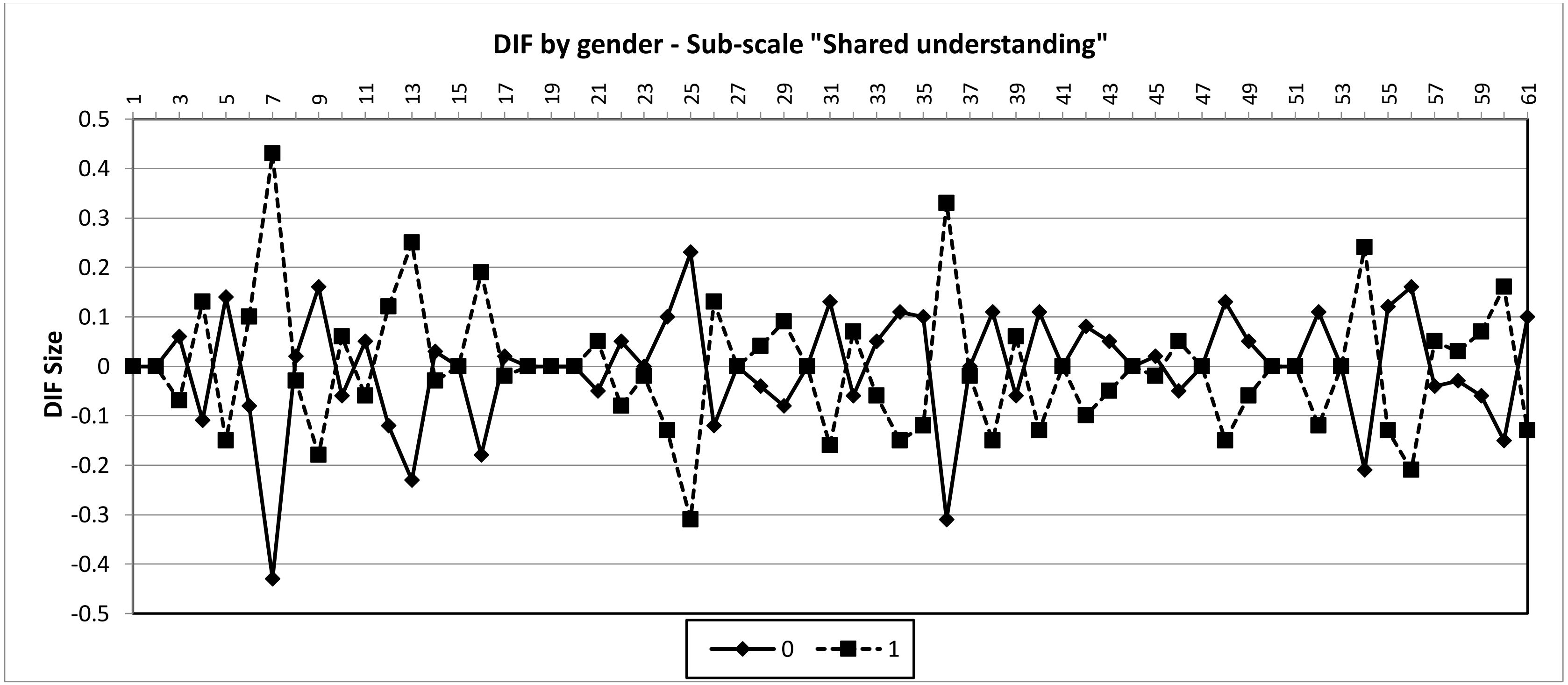


Figure 10.2. Person fit plot by gender (0 = male, 1 = female) – Sub-scale "Shared understanding" with 61 items

Notes: Items in the figure are placed by item order (from 1 to 61). To see the item code that they refer to see item fit statistics table (Table 10.17).

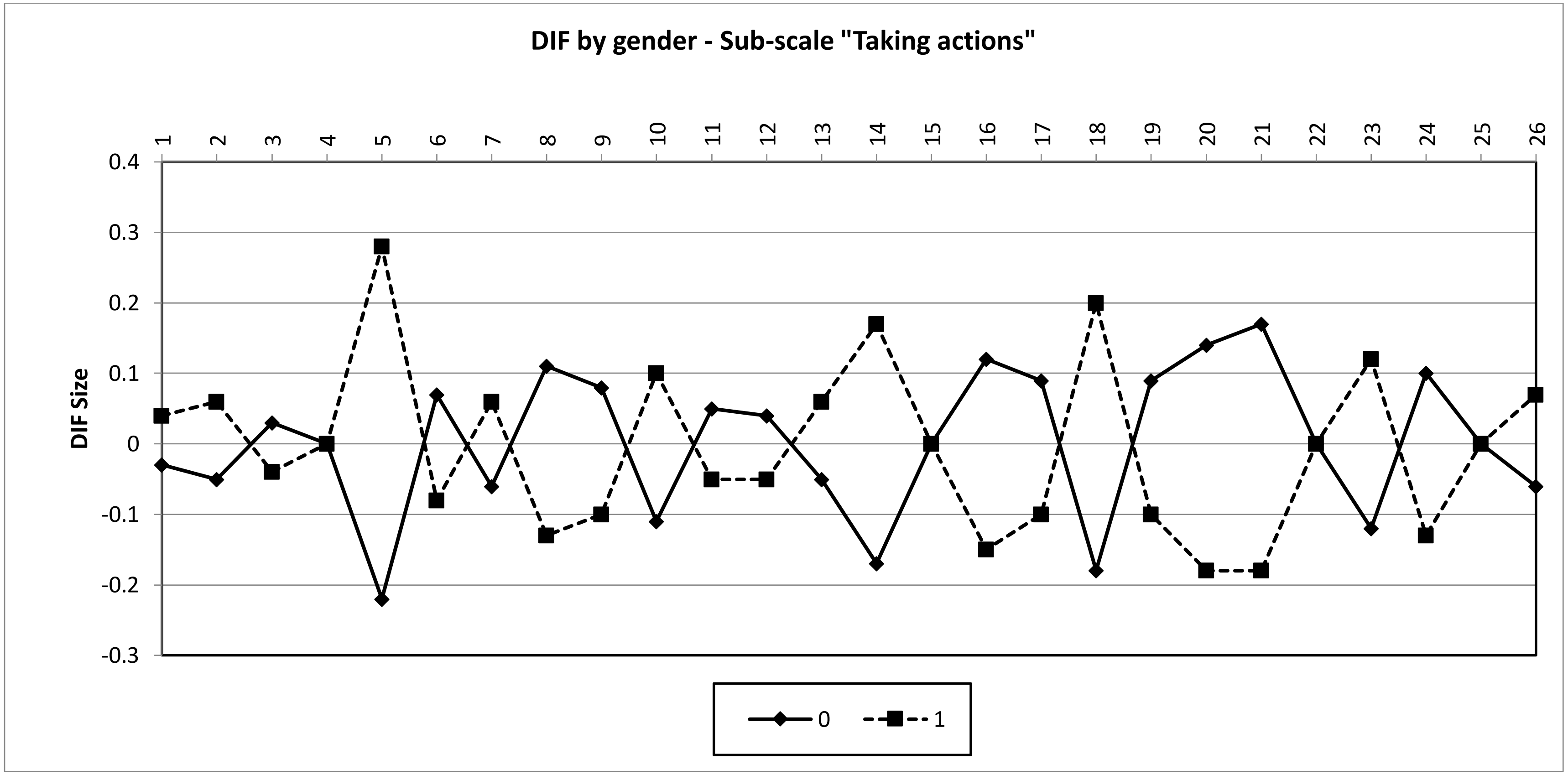


Figure 10.3. Person fit plot by gender (0 = male, 1 = female) – Sub-scale "Taking actions" with 26 items

Notes: Items in the figure are placed by item order (from 1 to 26). To see the item code that they refer to see item fit statistics table (Table 10.18).

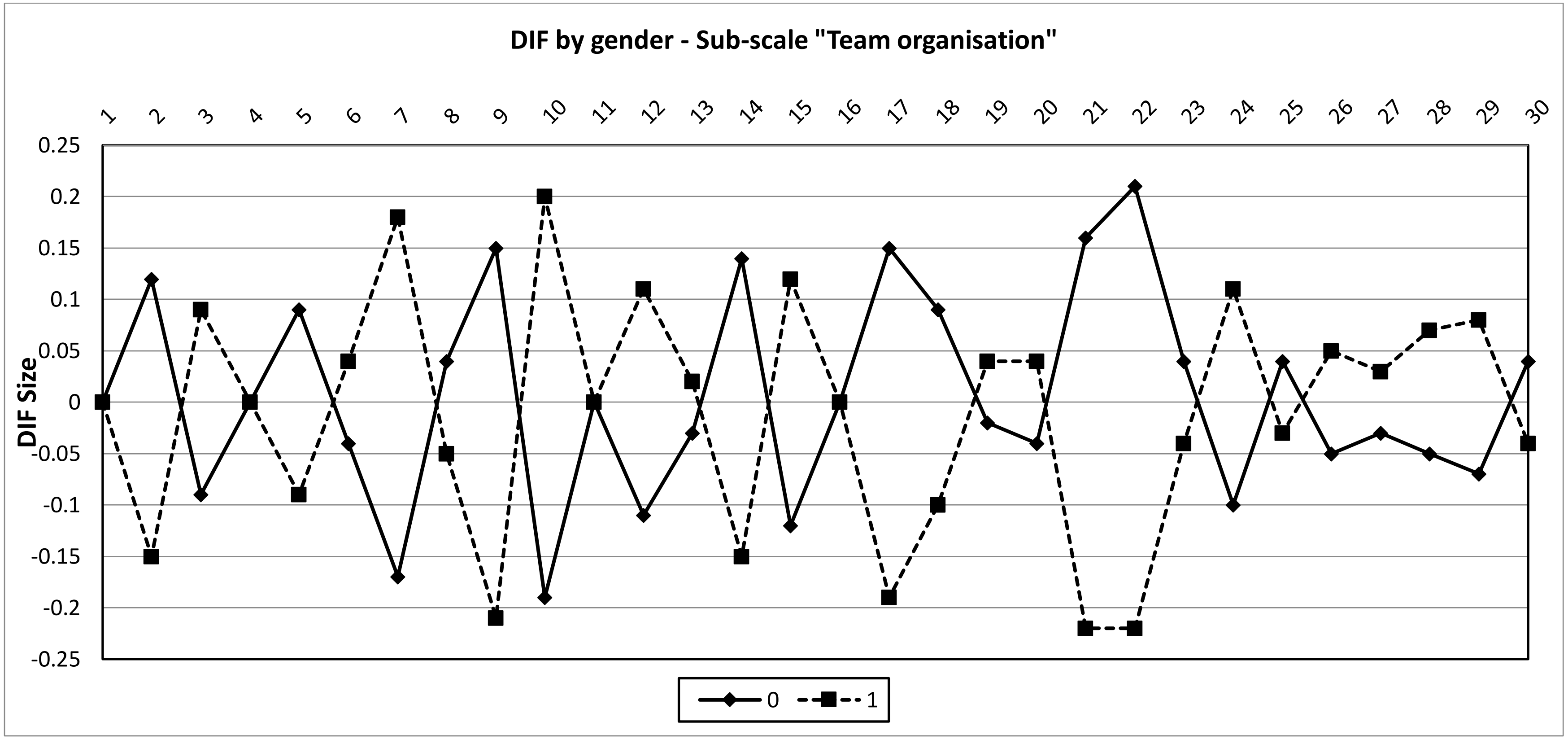


Figure 10.4. Person fit plot by gender (0 = male, 1 = female) – Sub-scale “Team organisation” with 30 items

Notes: Items in the figure are placed by item order (from 1 to 30). To see the item code that they refer to see item fit statistics table (Table 10.19).

10.20 Appendix 20: Principal components analysis of residuals results

Table 10.20. Standardised residual variance (in Eigenvalue units) for Sub-scale 1: Shared understanding (61 items)

	Empirical			Modeled
Total raw variance in observations	87.5	100.0%		100.0%
Raw variance explained by measures	26.5	30.3%		31.2%
Raw variance explained by persons	15.8	18.0%		18.6%
Raw Variance explained by items	10.7	12.2%		12.6%
Raw unexplained variance (total)	61.0	69.7%	100.0%	68.8%
Unexplained variance in 1st contrast	2.1	2.4%	3.4%	
Unexplained variance in 2nd contrast	1.6	1.9%	2.7%	
Unexplained variance in 3rd contrast	1.6	1.8%	2.6%	
Unexplained variance in 4th contrast	1.5	1.7%	2.4%	
Unexplained variance in 5th contrast	1.4	1.6%	2.3%	

Table 10.21 Standardised residual variance (in Eigenvalue units) for Sub-scale 2: Taking actions (26 items)

	Empirical			Modeled
Total raw variance in observations	47.2	100.0%		100.0%
Raw variance explained by measures	21.2	44.9%		44.7%
Raw variance explained by persons	8.4	17.8%		17.7%
Raw Variance explained by items	12.8	27.1%		26.9%
Raw unexplained variance (total)	26.0	55.1%	100.0%	55.3%
Unexplained variance in 1st contrast	1.4	3.0%	5.4%	
Unexplained variance in 2nd contrast	1.4	3.0%	5.4%	
Unexplained variance in 3rd contrast	1.3	2.7%	4.9%	
Unexplained variance in 4th contrast	1.3	2.7%	4.9%	
Unexplained variance in 5th contrast	1.2	2.6%	4.8%	

Table 10.22 Standardised residual variance (in Eigenvalue units) for Sub-scale 3: Team organisation (30 items)

	Empirical			Modeled
Total raw variance in observations	46.7	100.0%		100.0%
Raw variance explained by measures	16.7	35.7%		35.3%
Raw variance explained by persons	7.9	16.9%		16.7%
Raw Variance explained by items	8.8	18.9%		18.6%
Raw unexplained variance (total)	30.0	64.3%	100.0%	64.7%
Unexplained variance in 1st contrast	1.4	3.1%	4.8%	
Unexplained variance in 2nd contrast	1.4	3.0%	4.7%	
Unexplained variance in 3rd contrast	1.3	2.8%	4.3%	
Unexplained variance in 4th contrast	1.3	2.7%	4.2%	

10.21 Appendix 21: Methodological review - Cognitive interviewing

A search of the literature across all disciplines using the academic database Scopus shows the growth in the use of CI. A total of 2,564 references were identified (February 2022) across all disciplines including the term “cognitive interview” or “cognitive interviewing” in their abstract, title or keywords. Figure 10.5 shows publications by year and Figure 10.6 shows publications by subject area. Scopus database includes ‘Education’ subject area under Social Sciences, and for that reason separate results are not presented for Education here.

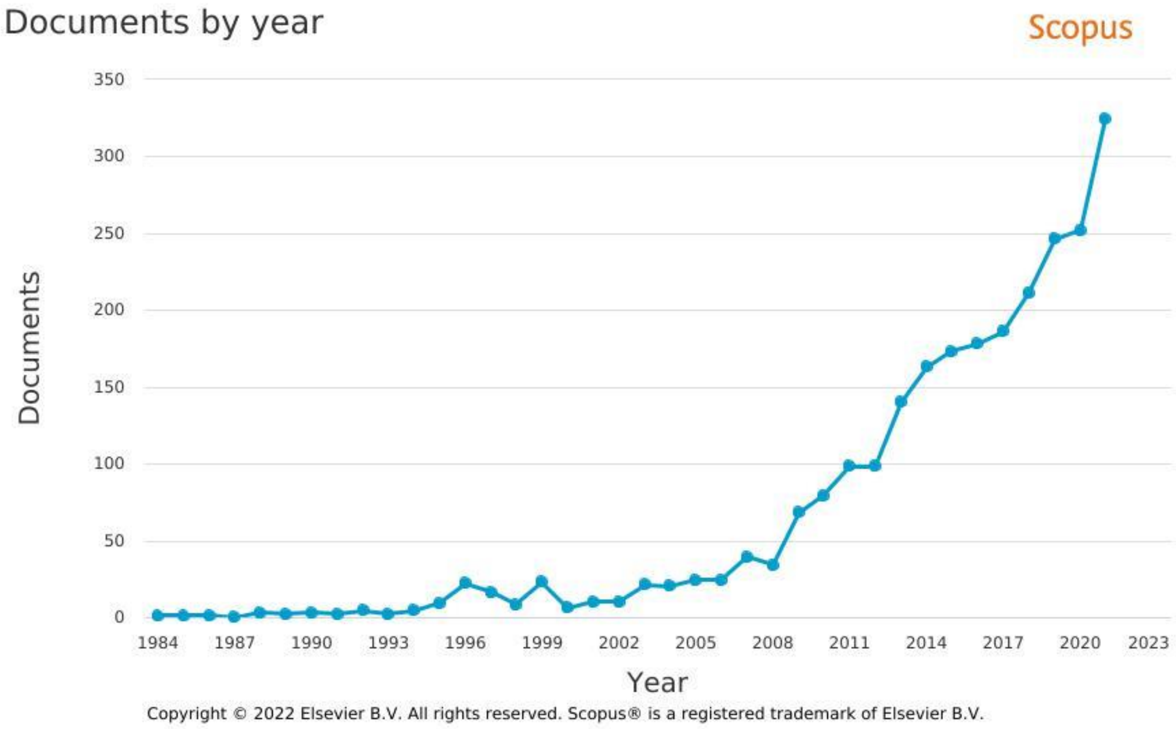


Figure 10.5. Publications by year – Cognitive interviewing search

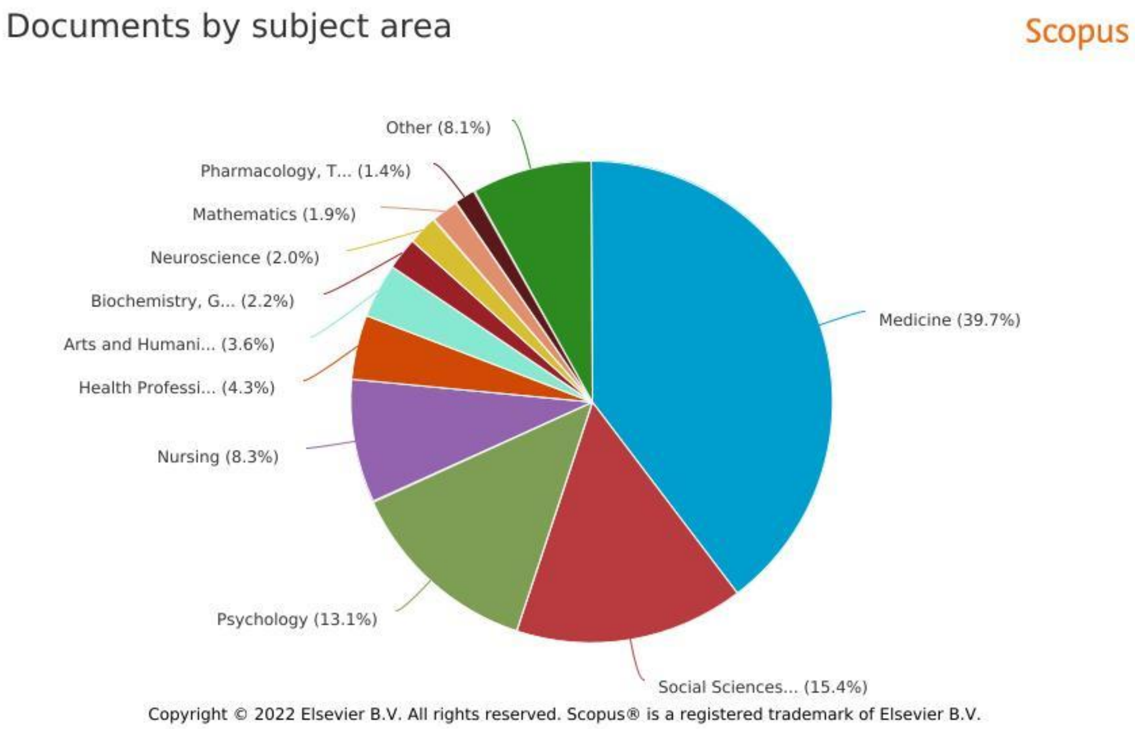


Figure 10.6. Publications by subject – Cognitive interviewing search