Collaborative Problem Solving:

Considerations for the National Assessment of Educational Progress

Collaborative Problem Solving:

Considerations for the National Assessment of Educational Progress

Authored by:
Stephen M. Fiore
Art Graesser
Samuel Greiff
Patrick Griffin
Brian Gong
Patrick Kyllonen
Christine Massey
Harry O'Neil
Jim Pellegrino
Robert Rothman
Helen Soulé
Alina von Davier

April 2017

Table of Contents

1 Executive Summary

Introduction

There is growing evidence from multiple sources within the United States as well as internationally that education is not preparing students for the workforce demands of today, much less tomorrow. While the acquisition of content knowledge remains critically important, it is not enough. The changing global economic and societal realities, as well as rapid technological transformations are reshaping life and work, as well as redefining and reprioritizing the skills that employees and citizens must have in order to succeed.

One skill that has attracted increased attention is *collaborative problem solving* (CPS). Increasingly, in a wide range of workplaces, employees work in teams—face-to-face and with peers around the country or around the globe—to develop solutions to non-routine problems. Data on the need for CPS competency come from numerous surveys, reports, and research studies over the past two decades.

Two international programs—the Assessment and Teaching of 21st Century Skills (ATC21S), a joint effort led by Cisco, Intel, and Microsoft; and the Programme for International Student Assessment (PISA), a 60-nation assessment administered by the Organization for Economic Cooperation and Development (OECD)—have developed assessments of collaborative problem solving. The National Center for Education Statistics (NCES) is considering adding such an assessment to the National Assessment of Educational Progress (NAEP), a federally sponsored assessment administered to a sample of students in a variety of subject areas.

To assist in its deliberations, NCES assembled an array of internationally recognized experts to examine state of the art research on collaborative problem solving and its assessment. The experts addressed the scope of the domain, as well as issues of assessment design, development, scoring, and reporting. The result of their work is this report. The purpose of this white paper is to inform decision making around the potential inclusion of CPS in NAEP. It is not intended to make recommendations about how NAEP should approach CPS, but rather to outline the possibilities for consideration.

Defining Collaborative Problem Solving

The term "collaboration" has different meanings in different environments. In K-12, collaboration almost always means an individual task can be solved by anyone in the group, but collaboration is also an instructional strategy to enable learning more efficiently or effectively. In the world of work (industry, military), the term "collaborative" usually means a group task in which no one member of the group can solve the task alone.

Collaborative problem solving involves two different constructs—collaboration and problem solving. The assumption is that collaboration for a group task is essential because some problem-solving tasks are too complex for an individual to work through alone or the solution will be improved from the joint capacities of a team. People vary in the information, expertise, and experiences that they can bring to bear in order to jointly solve a particular problem. More specifically, collaborative problem solving requires that people share their resources and their strategies in order to reach a common goal via some sort of communication process. Whether in an individual or group task, the group can be either face-to-face or virtual. In both cases, some technology is often used to facilitate collaborative problem solving.

Generally, collaborative problem solving has two main areas: the collaborative (e.g., communication or social aspects) and the knowledge or cognitive aspects (e.g., domain-specific problem-solving strategies). These two areas are often referred to as "teamwork" and "taskwork." The primary distinction between individual problem solving and collaborative problem solving is the social component in the context of a group task. This is composed of processes such as the need for communication, the exchange of ideas, and shared identification of the problem and its elements.

The PISA 2015 framework defines CPS as follows:

> Collaborative problem solving competency is the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution.

Competency is assessed by how well the individual interacts with agents during the course of problem solving. This includes achieving a shared understanding of the goals and activities, as well as efforts to pool resources and solve the problem.

Within the PISA framework, three competencies form the core of the collaboration dimension: establishing and maintaining shared understanding, taking appropriate action to solve the problem, and establishing and maintaining group organization. The framework also identifies four problem-solving processes: exploring and understanding, representing and formulating, planning and executing, and monitoring and reflecting.

ATC21S has also developed a framework for assessing collaborative problem solving. This framework, like that of PISA, identifies dimensions of collaboration: participation, perspective-taking, and social regulation. Problem-solving skills include task regulation skills and knowledge-building and learning skills. The main distinction between the two frameworks is that ATC21S has an integrated approach, in which the distinctions between collaboration and problem solving are melded. For the purposes of defining collaborative problem solving, research must determine which of these approaches produces the appropriate level of granularity and accuracy in assessment.

Other research on cognition has identified some additional competencies that could be measured in an assessment of collaborative problem solving. These include group communication, "macrocognition" in teams, collaborative learning, and cognitive readiness.

Assessment Design

In addition to the definition of collaborative problem solving, there are a number of issues associated with the design of an assessment of that competency. These include the following:

- What is the nature of the assessment task(s)? What content is to be assessed? Is the task ill-defined or well-defined?
- Who is collaborating? Are they human-to-human or human-to-agent interactions? How many members are on the teams?
- What is the composition of the team? Is it homogeneous or heterogeneous? Do the team members assume similar or different roles?
- What is being evaluated? Is it individual or team outcomes? Is it process variables?

Assessment Development

The experience of ATC21S and PISA suggests that assessment task development for collaborative problem solving differs little from standard test development. The major departure is in conceptualizing what constitutes a test item and how students will respond to it. The ATC21S assessment involves human-to-human interaction, while the PISA assessment involves human-to-agent interaction. The choice of the approach has strong implications for many of the other issues. For instance, what will the task environment look like? What will be the group composition? What collaboration medium should be used? How can the scoring be implemented?

The human-to-human assessment approach is embedded in a less standardized assessment environment and offers a high level of face validity. A student collaborates with other students, so the behavior of both is difficult to control. Also, the success of one student depends on the behavior of the other student, as well as the stimuli and reactions that he/she offers. This has implications for scoring. How can the open conversation and large variety of stimuli be identified and utilized for scoring?

In the human-to-agent approach, the assessment environment is more standardized. The behavior of computer agents must be preprogrammed, so the response alternatives of the item to which the student reacts need to be limited to some extent and every possible response of the student needs to be attached to a specific response by the computer agents' stimuli or event in the problem scenario. Such an approach ensures that the situation is rather standardized, and furthermore enables comprehensive scoring techniques since every possible turn in collaboration is predefined. Nevertheless, the human-to-agent approach has the shortcoming of its artificial appearance and other conceptual questions, such as whether the CPS measured in this way can represent real-life collaborative problem solving.

Scoring Collaborative Problem-Solving Assessments

Data from CPS tasks can be characterized as individual and team (collective) outcome data and process data. Examples of outcome data are the correct/incorrect assessment of an action or task at the individual or team level. The process data from CPS consist of time-stamped sequences of events registered in a log file which describe the actions and interactions of the students and the system.

The outcome data are collected through evaluative scoring throughout the process of collaboration. For example, an individual's actions during the collaboration can be scored as *correct* or *incorrect* by a human rater or an automatic scoring engine. The team-level outcome data are straightforward to collect. These data indicate whether a team solved the problem successfully.

There may be some process variables that are relatively easy to measure, such as the participation level of each team member and turn-taking. However, beyond these kinds of variables, interpreting actions and chats may be much more complex because of the dynamics and the sheer volume and complexity of data generated in log files.

Scoring collaborative problem-solving tasks is still very much in the research phase. More needs to be done before we can confidently make inferences about collaborative skills based on item responses. Outcome data (correct/incorrect/partially correct) can be analyzed with classical test theory (reliability, correlations, biserial correlations), factor analysis, and item response theory (IRT). Process data may also be analyzed this way under specific circumstances and may also be analyzed using data mining tools and time series methods. It is important to note that the scoring model used is dependent on prior decisions about what should be measured and how. Assessment design and development must be undertaken in a way that aligns with the intended scoring model.

Reporting Results

Reporting is the most publicly visible component of a NAEP assessment. What is reported from an assessment plays a major role in how the assessment is received both by the general public and by the more direct stakeholders, including teachers and school, district, and state personnel.

There are a number of issues NCES must address in determining how to report results from a collaborative problem-solving assessment, including

- What will be reported: collaborative skill, problem-solving skill, or both? Will it be individual responses or team responses? Will it be within a content area or separately?
- What is the score structure: one scale or separate scales for sub-skills? Will the scores be reported using performance levels? Will the scores be mapped to sample items?

NAEP and PISA provide excellent models for score reporting on national and international assessments. Regular NAEP assessments, such as mathematics and reading, provide good models for design, analysis, scoring, and reporting using probability (cluster) sampling, balanced incomplete block designs for administration, item-response theory for scoring, scaling, and linking enabling comparisons across grades and across time. Unique NAEP items, such as the science hands-on tasks and interactive computer tasks, and the new Technology and Engineering Literacy (TEL) tasks provide other models for reporting. PISA 2015, the score report that was released in December 2016, provides a reporting model for collaborative problem solving.

Implications for NAEP

NAEP's importance and credibility has provided the assessment with a large influence on policy and practice at the state and local levels. Because of NAEP's prominence, NAEP and the National Assessment Governing Board need to exert caution in developing and implementing a new type of assessment. An assessment of collaborative problem solving would introduce a number of new elements to NAEP that can pose challenges to interpretation. These include the use of multiple content areas (and possibly content-neutral problems), the use of student groups in test administrations, and the use of process variables. NCES and the Governing Board should be very clear about the information the assessment results provide, as well as their limitations.

Despite these challenges, an assessment of collaborative problem solving would represent a bold step that could strengthen NAEP's role as an assessment leader. As a so-called "21st century skill," collaborative problem solving is considered vital to students' future in the workplace. Policymakers and practitioners need to know how well students can demonstrate that skill, and NAEP is in an ideal position to provide that information.

2 Introduction

The Need for Collaborative Problem-Solving Assessments

There is an increased interest in collaboration and teamwork in the workforce, higher education, and K-12 education. The Assessment and Teaching of the 21st Century Skills (ATC21S) included collaboration among the most important skills necessary for a successful career. While there is widespread agreement in the field of education that collaboration is an important skill, there is less agreement on how to build an assessment to measure it.

In 2015 the Organization for Economic Cooperation and Development (OECD) published its draft framework, which included strong rationale for the inclusion of CPS, calling collaborative problem solving "a critical and necessary skill across educational settings and in the workforce." This plan by the OECD to add CPS to its 2015 PISA test has important implications for U.S. assessments, including the National Assessment of Educational Progress (NAEP).

NAEP, which is the largest nationally representative assessment that is administered regularly over time, measures what America's students know and can do in various subject areas. A broad range of audiences use the assessment results, including policymakers, educators, and the general public. Subject areas range from traditional curricular subjects, such as mathematics, to non-curricular topics, such as technology and engineering literacy. Each NAEP assessment is built around an organizing conceptual framework. Assessments must remain flexible to mirror changes in educational objectives and curricula; hence, the frameworks must be forward-looking and responsive, balancing current teaching practices with research findings. The most recent NAEP framework created was the Technology and Engineering Literacy (TEL) framework, which was introduced in 2014.

Because of the growing importance of collaborative problem-solving skills in the educational landscape, the National Center for Education Statistics (NCES) decided that NAEP should investigate state-of-the-art CPS research and assessment before deciding whether an assessment of CPS should be added to NAEP. Therefore, NCES assembled a broad array of individuals to develop this white paper, with the goal of fully conceptualizing the assessment of CPS skills as it currently exists,

to inform not only NCES, but also the broader field of researchers and policymakers interested in CPS. This paper represents the culmination of a process that began with the NAEP Symposium on Assessing CPS Skills in September 2014.

In this white paper, the CPS assessment is assumed to be part of the NAEP context in terms of the data collection design, instruments, and sampling. That is, it is assumed that a NAEP CPS test would be a group-score assessment as opposed to an individual-score test (Mazzeo, Lazer, & Zieky, 2006); not tied to a particular curriculum; permitting the use of item pools, rather than fixed test forms; and administered in a matrix-sampling design (i.e., a student only takes a subset of the item pool). It is also assumed that summary scores (e.g., average scores, percentage achieving a certain level of proficiency) would be reported at the national, state, and/or jurisdiction level but not at the individual student level, and that scores would be disaggregated by major reporting subgroups, including gender (male-female), race/ethnicity, socioeconomic status levels (e.g., full, partial, or not eligible for the National School Lunch Program), English language learner (ELL) status, and students with disabilities (SD) status. It is also assumed that the assessment would be administered at the fourth-, eighth-, and twelfth-grade levels. As with all NAEP assessments, the assumption is that (a) the resulting CPS scores should be accompanied by statements about the precision of the measurements, and (b) scores are valid and fair for all major test takers groups (see AERA, APA, NCME Standards, 2014; Thissen & Wainer, 2001).

The rest of this introductory chapter is organized into two sections. The first section discusses the value of CPS as a 21st century skill, its importance to education and workforce readiness, and the need for CPS assessments. The second section briefly reviews the history of the research around collaboration and collaborative problem-solving constructs, key questions to be addressed, and the challenges associated with assessing CPS.

COLLABORATIVE PROBLEM SOLVING: AN ESSENTIAL 21ST CENTURY SKILL

There is growing evidence from multiple sources within the United States as well as internationally that education is not preparing students for the workforce demands of today, much less tomorrow. While the acquisition of content knowledge remains critically important, it is not enough. The changing global economic and societal realities, as well as the rapid technological transformations, are reshaping life and work in addition to redefining and reprioritizing the skills that employees and citizens must have in order to succeed.

Beginning in the early 1990s and continuing to the present, numerous reports, surveys, and research have captured the data on the transformation of work and subsequent changing requirements for the workforce. The U.S. Department of Labor's Secretary's Commission on Achieving Necessary Skills (SCANS) report in 1991, *What Work Requires of Schools*, first documented this as a coming problem (U.S. Department of Labor, 1991). The second report, *Skills and Tasks for Jobs*

issued in 2000, sought "to help educators make high school courses more relevant to the needs of a modern workforce and to help employers ensure that their employees possess appropriate, up-to-date skills" (U.S. Department of Labor, 2000). Subsequent surveys such as *Are they Ready to Work* (2006) by The Conference Board, Inc., the Partnership for 21st Century Skills, Corporate Voices for Working Families, and the Society for Human Resource Management have added business' perspective regarding the new sets of knowledge and skills needed for the 21st century U.S. workforce. Key research findings such as the work of Levy and Murnane (2013) in *Trends in Routine and Non-routine Tasks in Occupations, United States, 1960 to 2009*, have cited the dramatic increase in demand for non-routine interpersonal and analytic skills and the corresponding decrease in demand for routine manual and cognitive skills.

For more than two decades, the data has consistently identified collaboration, critical thinking, problem solving, communication, and creativity/innovation as critically important skill sets for the future. The 2010 American Management Association P21 Critical Skills Survey of 2000+ business leaders found that over 70% of leaders surveyed identified these skills as priorities when hiring and evaluating employees. In March 2015, the World Economic Forum released a report, *New Vision for Education*, based on analysis of research from nearly 100 countries, and again reiterated the ongoing skills gap and the importance of 21st century skills, including collaboration and problem solving: "These gaps are clear signs that too many students are not getting the education they need to prosper in the 21st century and countries are not finding enough of the skilled workers they need to compete" (2015).

By 2002, organizations such as the Partnership for 21st Century Skills (P21), a non-profit coalition of business, education, and community leaders, began to emerge to build awareness of the seriousness of the skills gap and advocate for educational solutions. P21, with the assistance of a broad coalition of experts from business, education and government, developed the Framework for 21st Century Learning, a comprehensive set of student outcomes that articulated necessary skills, content knowledge, and interdisciplinary themes needed for the future. Central to the framework are the 4 C's—collaboration, critical thinking and problem solving, creativity and innovation, and communication—the very skills that business, education, and researchers have identified as critical for the future. The framework has been documented in numerous books and most models of 21st century skills include collaboration, or collaborative problem solving, as important skills. (e.g., Fadel & Trilling, 2012; Trilling & Fadel, 2009; Wagner, 2008, 2010).

The National Research Council's report *Education for Life and Work, Developing Transferable Knowledge and Skills in the 21st Century* (July 2012) analyzed the models and research (or lack thereof) around "21st century skills" and "deeper learning." The report articulated the importance of skills and proposed clarifying the terms by dividing them into three domains: intrapersonal, interpersonal, and cognitive. Under this framework, collaboration resides in the interpersonal domain and problem solving in the cognitive domain. Additionally, the report outlined educational strategies and significant research recommendations.

The importance of collaborative problem solving as an educational outcome and important skill for life and work has continued to increase since the turn of the 21st century. Business continues to evolve requiring more cross-functional teams that work across international and cultural borders and possess complex cognitive, collaborative, and critical thinking skills (American Management Association, 2010). As Dede (2009) has observed,

> The nature of collaboration is shifting to a more sophisticated skillset. In addition to collaborating face-to-face with colleagues across a conference table, 21st century workers increasingly accomplish tasks through mediated interactions with peers halfway across the world whom they may never meet face-to-face. … Collaboration is worthy of inclusion as a 21st century skill because the importance of cooperative interpersonal capabilities is higher and the skills involved are more sophisticated than in the prior industrial era.

It is no secret that traditionally the U.S. K-12 education system has focused almost solely on content delivery. "Little time is spent on building capabilities in group interpretation, negotiation of shared meaning, and co-construction of problem resolutions" (Dede, 2009). Developing 21st century skills, especially the 4 Cs, has been left to students to learn on their own. The growing body of evidence, including student performance on national and international tests such as PISA, suggests that this is no longer acceptable. The incorporation of collaborative problem solving into the Common Core State Standards has brought additional focus to the importance of teaching and assessing it.

COLLABORATIVE PROBLEM SOLVING: A NEW ASSESSMENT DOMAIN

Collaborative problem solving is a new assessment domain which is developing rapidly; however, there is much to be learned. The Assessment and Teaching of 21st Century Skills (ATC21S) CPS assessment has emerged out of a coalition of advisors and experts as part of a project sponsored by Microsoft, Intel, and Cisco, and part of a larger effort to define, develop pedagogies for, and assess 21st century skills (Griffin, 2014). Like PISA, tasks are designed to elicit collaborative problem-solving behaviors by having students work in pairs and communicate through on-screen chat messaging. Unlike PISA, students are collaborating with other students rather than with computer agents. There is a clear delineation between social skills and cognitive skills. The ATC21S model emphasizes the development of both collaborative skills and the cognitive outcomes that can result from student mastery of those collaborative skills.

As mentioned earlier, OECD's addition of CPS to its 2015 PISA test has important implications for U.S. assessments including NAEP (OECD, 2013). Internationally, significant focus on these skills has emerged in the educational plans of many countries – for example, Singapore's ICT Masterplan and Israel's national plan for Adapting the Educational System to the 21st Century. PISA's inclusion of collaborative problem solving in its assessment is also driven by the need to teach and assess the skills most in demand, not just in the U.S., but internationally as well (Brannick & Prince, 1997; Griffin, 2014; National Research Council, 2011; Rosen & Rimor, 2012).

The PISA 2015 CPS competency is a conjoint dimension of collaboration skills and the skills needed to solve the problem (i.e., referential problem-solving skills), while collaboration serves as a leading strand. The CPS framework for PISA incorporates definitions and theoretical constructs that are based on research and best practices from several areas where CPS-related skills have been assessed. The PISA 2015 CPS framework further incorporates information from existing assessments and frameworks including ATC21S, PIAAC , the P21 Framework for 21st Century Learning, the Center for Research on Evaluation, Standards, and Student Testing's (CRESST) six measures, and Collazos' five system-based indicators of CPS success, to name a few.

Work continues in this area. In November 2014, Educational Testing Service (ETS) and the Army Research Institute hosted a two-day working meeting on "Innovative Assessments of Collaboration," and a new edited volume on the assessment of CPS skills is in progress (von Davier, Zhu, & Kyllonen, 2017). A research team at ETS has also developed a prototype for science collaborative assessment, the Tetralogue, with the goal of jointly assessing cognitive (science) skills and social (CPS) skills.

Collaborative Problem-Solving Skills: An Interdisciplinary Experience

As discussed above, the rapidly growing literature on teamwork and collaborative problem solving suggests that these skills are becoming increasingly more important to today's organizations. Teamwork and collaborative problem solving are among the most frequently mentioned 21st century skills (National Academies, 2012; Wildman et al., 2012; Casner-Lotto & Barrington, 2006). Employers and policymakers point to the importance of teamwork and collaboration skills in the current and future workplace, and educational systems are starting to recognize the importance of teaching those skills, for example, through authentic, "real world" experiences (ATC21S; Mazur, 1997).

However, challenges persist. Significant developments in these domains often occur across different communities which are largely independent of one another, and CPS skills are among those skills that are considered hard-to-measure (Stecher & Hamilton, 2014).

In education, collaborative problem solving is an emerging topic with its own literature focusing on student learning processes and assessment. In psychology, the analysis of team skills typically focuses on the dynamics and coordination of team members in various types of teams. In engineering, the focus sits at the intersection of understanding of the collaborative processes and design of interactive systems to assist such processes. To date, there is little knowledge sharing across these communities.

The second major challenge across all the communities concerns the assessment of collaborative skills. Traditional methods—item response and classical test theories—may not be appropriate for measuring collaborative interactions because of the dependence within elements of complex tasks and between interacting participants (Cooke et al., 2012; Quellmalz et al., 2009). New assessment designs and statistical methods that capture the dynamic of knowledge sharing in collaborative

contexts may be required (Dede, 2012). The main challenge of the next generation of assessments, and the next generation of psychometrics, is how to model this knowledge in a way that meets the technical standards of traditional assessments.

In order to address these issues, the team of contributors to this white paper is diverse. To ensure that our interdisciplinary contributors produce a relatively integrated view, the authors address the following key questions:

1. How can we define collaboration and team output? What are the features of a good team?
2. How can we define the contributions of an individual to team performance?
3. How do we design an assessment that captures the process of collaboration together with the outcome of collaboration?
4. What are the features of a good interactive CPS task? Should it be a simulation or a game-based task? Is human-to-agent collaboration a good approximation of the CPS skills needed for the human-to-human interaction?
5. What type of interdependencies exist in the data collected from interactive CPS tasks? What is the best way to capture and analyze them? Will interactive CPS tasks result in "Big Data," and if so, what is the implication for large assessments like NAEP?
6. What kind of statistical methods can be used to evaluate team dynamics? What is the unit of observation and the scoring rubric (Nelson, 2007)? For example, should we consider the specific actions taken by each student (and captured in log files)? Or should interim but summative stages for scoring be considered for an entire team? Should we model individual ability in the area of the assessment or the collective ability of a team?
7. Are there methodologies in use in one domain (e.g., group dynamics) that could be applied to another (e.g., collaborative problem solving)?
8. How do we score a CPS task? Should we report a multidimensional score or a combination of an individual score and a team score? Should we consider automatic scoring engines for analyzing the discourse data?
9. How will we define fairness in a CPS task?

These questions pinpoint features of the CPS construct that make its measurement considerations significantly different from traditional cognitive or even non-cognitive tests. Each of these issues is discussed in the individual chapters.

The purpose of this white paper is to inform decision making around the potential inclusion of CPS in NAEP. It is not intended to make recommendations about how NAEP should approach CPS, but rather to outline the possibilities for consideration.

3 Defining Collaborative Problem Solving (CPS)

Introduction

Collaborative problem solving (CPS) is composed of two main elements: the collaborative, sharing, or social aspects coupled with the knowledge or cognitive aspects. Thus, the primary distinction between individual problem solving and collaborative problem solving is the social component. This is comprised of processes such as the need for communication, the exchange of ideas, shared identification of the problem, negotiated agreements, and relationship management. It is important to distinguish CPS from other forms of collaboration. In collaborative problem solving there is a group goal that needs to be achieved, the solution requires problem solving, team members contribute to the solution, and there is some foundation for evaluating whether the group goal has been achieved. Moreover, the activities of the team members are interdependent, with various roles, so that a single person cannot solve the group goal alone. The collaborative activities therefore require communication, coordination, and cooperation. These characteristics of CPS differ from collaborative learning, where the emphasis is on the learning of subject matter by individual team members in a group context but does not require a group problem that has to be solved with interdependency among team members. Similarly, collaborative work and collaborative decision making involve interdependent joint activities, but do not require problem solving.

This chapter defines collaborative problem solving that involves multiple people working interdependently towards a common goal (Fiore, 2008; Fiore et al., 2010; Graesser, Foltz et al., in press; Griffin, 2014). We describe two major frameworks that have been developed to define collaborative problem solving for the purposes of assessment. The first is the Program for International Student Assessment (PISA) assessment (Graesser et al., in press; OECD, 2013), which assessed CPS in several dozen countries in 2015. The second is the Assessing and Teaching 21st Century Skills (ATC21S) project, headquartered at the University of Melbourne, that developed a conceptual framework for CPS in 2010 with associated assessment tasks (Griffin, McGraw, & Care, (Eds.) 2012; Hesse, Care, Buder, Sassenberg, & Griffin, 2015). We then discuss some approaches to CPS that include additional factors that might need to be considered in CPS assessment.

An adequate understanding of CPS and its development requires a definition of the construct as well as a way of measuring it. This chapter defines the various taskwork and teamwork dimensions associated with CPS whereas subsequent chapters address the measurement of the construct. In collaborative problem solving, taskwork refers to the activities associated with accomplishing an objective (e.g. solving a problem), while teamwork refers to the activities associated with collaboration (e.g. communicating with teammates, delegation of responsibilities, resolving conflict).

PISA (2015) Collaborative Problem-Solving Framework

The Organization for Economic Cooperation and Development (OECD) examines educational performance in terms of the skills that students acquire rather than the number of years of formal education completed. It does this through its Program for International Student Assessment (PISA). It has also engaged in similar activities through its International Adult Literacy Surveys, its Program for the International Assessment of Adult Competencies (PIAAC), and its forthcoming Assessment of Higher Education Learning Outcomes (AHELO). In 2015 collaborative problem solving was included in the PISA assessments for the first time.

Definitions of CPS all assume that a group includes two or more individuals whose task is to solve a problem that cannot be successfully or efficiently completed by only one individual. The PISA 2015 framework defined CPS as follows:

> Collaborative problem-solving competency is the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution (OECD, 2013, p. 6).

Competency is assessed by how well the individual interacts with agents during the course of problem solving. This includes achieving a shared understanding of the goals and activities, as well as efforts to pool resources and solve the problem.

The word *agent* refers to either a human or a computer-simulated participant. In both cases, an agent has the capability to generate goals, perform actions, communicate messages, react to messages from other participants, sense its environment, and adapt to changing environments. Computer agents were selected in the PISA 2015 for a variety of logistical reasons. Specifically, there were limited time constraints in the assessment (two 30-minute sessions) and the individual needed to experience a diversity of groups, team members, tasks, and challenges in order to achieve a reliable and valid assessment of the construct. In addition, group members sending text messages to each other to solve the problem was not feasible for PISA due to the limits of networking capacity, scheduling protocols, and the ability of the computer to interpret the messages in an assessment that involves several dozen languages, countries, and cultures.

COLLABORATION AND PROBLEM-SOLVING PROCESSES

The PISA framework organized CPS competencies along two dimensions in a matrix that is shown in Table 3.1. One dimension addresses collaboration (the columns with 3 components) and the other problem solving (the rows with 4 components). The resulting matrix has 12 cells with skills that integrate collaboration and problem-solving processes. The collaboration dimension has three competencies. First is the concept of "establishing and maintaining shared understanding." Here, students need to identify shared knowledge (what each other knows about the problem), identify the perspectives of other agents in the collaboration, establish a shared vision of the problem states and activities, and monitor and maintain relevant shared knowledge throughout the problem-solving task. Concrete actions include responding to requests for information, sending important information to agents about whether tasks are completed, verifying what each other knows, negotiating agreements, and repairing communication breakdowns. Second is the idea of "taking appropriate action to solve the problem." Here, students need to identify the type of CPS activities that are needed to solve the problem and to follow the appropriate steps to achieve a solution. Both physical actions and acts of communication are typically needed to solve the problem. Third is the idea of "establishing and maintaining team organization." Students need to organize and monitor the team to solve the problem; consider the talents, resources, and assets of team members; understand the roles of the different agents; follow the relevant steps for assigned roles; and reflect on the success of the team organization.

The problem-solving dimension directly incorporates the same competencies that were adopted in the individual problem-solving assessments of the PISA 2012 framework (Funke, 2010; Greiff et al., 2013; OECD, 2010). Specifically, there are four problem-solving processes that are needed to successfully address the kinds of complex problems being faced in the 21st century. First is the idea of "exploring and understanding." In this stage, students need to interpret initial information about the problem, which is uncovered during exploration and interactions with the problem. Second is the notion of "representing and formulating." Here students must select, organize, and integrate information with prior knowledge. From a process standpoint, this may include the use of graphs, tables, symbols, or words. Third is the process of "planning and executing." This includes identifying the goals of the problem, setting any sub-goals, developing a plan to reach the goal state, and executing the plan. Last is the phase of "monitoring and reflecting." Here the student is expected to monitor steps in the plan to reach the goal state, mark progress, reflect on the quality of the solutions, and revise the plan when encountering obstacles.

Table 3.1.

Matrix of collaborative problem solving for PISA 2015 (OECD, 2013)

	(1) Establishing and maintaining shared understanding	(2) Taking appropriate action to solve the problem	(3) Establishing and maintaining team organization
(A) Exploring and Understanding	(A1) Discovering perspectives and abilities of team members	(A2) Discovering the type of collaborative interaction to solve the problem, along with goals	(A3) Understanding roles to solve problem
(B) Representing and Formulating	(B1) Building a shared representation and negotiating the meaning of the problem (common ground)	(B2) Identifying and describing tasks to be completed	(B3) Describe roles and team organization (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 organization and roles

The 12-cell matrix provides a set of definitions for guiding assessment of collaboration skills engaged during problem solving. Some skills are assessed by the actions that the student performs, such as making a decision, choosing an item on the screen, selecting values of parameters in a simulation, or preparing a requested report. Other skills require acts of communication, such as asking other group members questions, answering questions, disseminating information, issuing requests, and giving feedback to others. Assessment of CPS, then, requires tracking the actions and communications of the individual being tested as the individual experiences a series of events, actions, and conversational speech acts by the other agents in the group. These experiences include obstacles to the collaboration, such as an agent that makes errors, fails to complete tasks, or does not communicate important information. The assessment tracks how well the human handles the various challenges and has skills associated with all 12 cells in the matrix. Details on the design, development, and scoring of items are provided in Chapters 4, 5, and 6.

COLLABORATIVE PROBLEM-SOLVING CONTEXTS

The PISA CPS 2015 framework also describes characteristics of problem-solving contexts that need to be considered. The problem-solving tasks require interdependency and joint activity among agents so that one agent cannot solve the group goal independently. Different problem-solving scenarios are identified, such as hidden profile (team members start out with distinct and complimentary pieces of information that must be shared), consensus in making a decision, and

negotiation. The roles of the team members are considered, such as whether they have different skills or power status. These considerations of problem-solving context constrained the selection of problem-solving tasks in the assessment.

Assessing and Teaching 21st Century Skills (ATC21S) Framework

The ATC21S conceptual framework for CPS guided the 2010 assessment in Australia (Griffin, McGraw, & Care, (Eds.) 2012; Hesse, Care, Buder, Sassenberg, & Griffin, 2015). CPS was viewed as a composite skill arising from the links between critical thinking, problem solving, decision making and collaboration. There was a distinction between inductive reasoning that focuses on establishing a possible explanation to test and deductive reasoning that involves testing whether the explanation is valid. It was argued that CPS requires developing skills to work both inductively and deductively in partnerships, to reach agreements on ideas, to form and test hypotheses, and to agree on strategies their team will use.

COLLABORATIVE PROBLEM-SOLVING MODEL

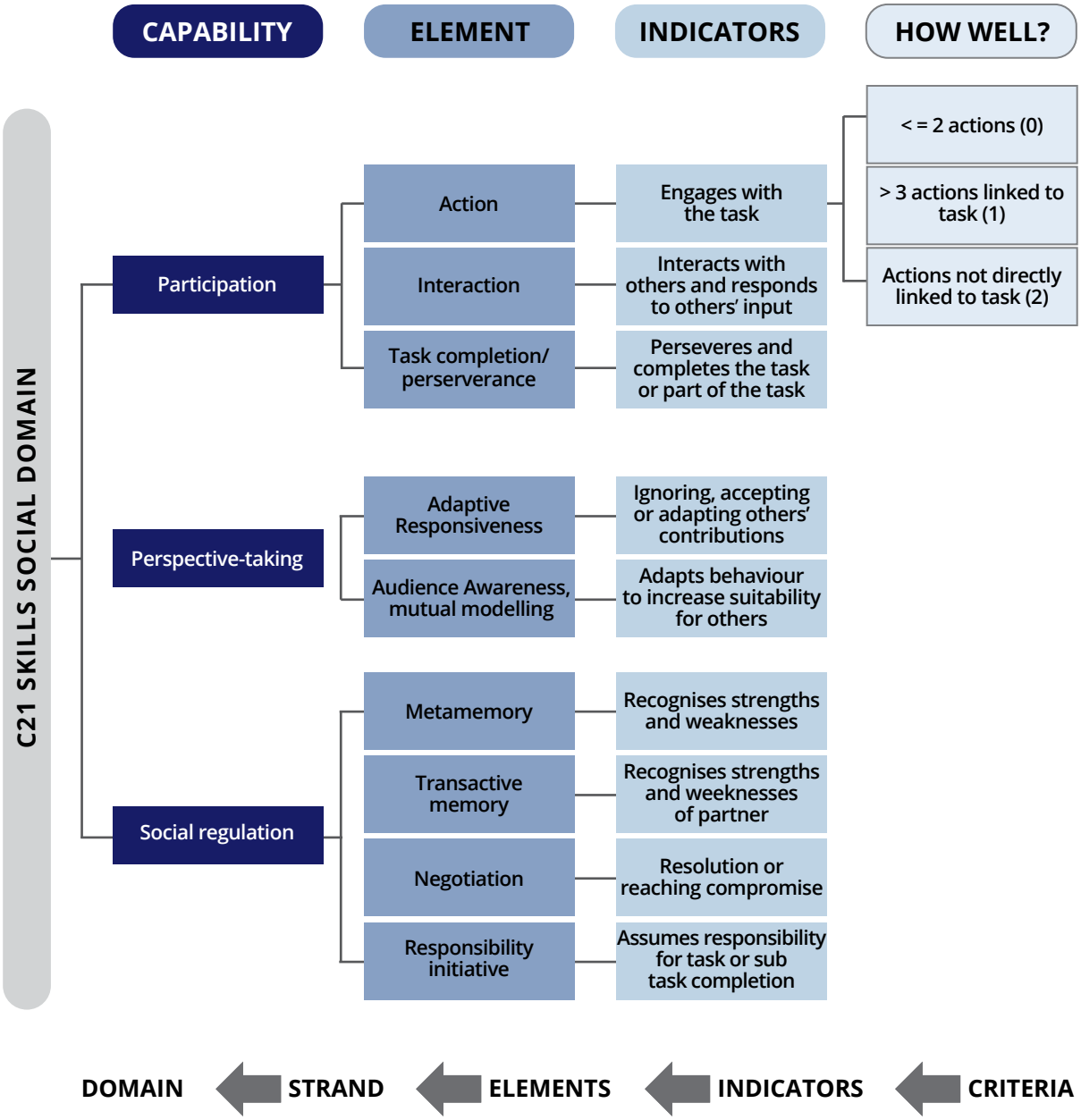
Social skills were identified in ATC21S in light of the fact that CPS requires a combination of collaboration (working together) and problem solving (using cognitive skills). Figure 3.1 shows the social skills reported in Hesse, Care, Buder, Sassenberg, & Griffin (2015) participation, perspective-taking, and social regulation. Participation is further broken down into action, interaction, and task completion. Perspective-taking involves responsiveness and audience awareness. Social regulation consists of negotiation skills, metamemory (the ability to evaluate one's own knowledge, strengths, and weaknesses), transactive memory (a person's understanding of the knowledge, strengths and weaknesses of collaborative partners), and responsibility initiative.

Each of the elements (second column) in Figure 3.1 is described by an example of an indicative, observable behavior (third column), and each behavioral indicator is further described by how much or how well the behavior is demonstrated (criteria, fourth column). For reasons of space, only one set of criteria is provided here, but in developing CPS tasks, attention had to be paid to a range of indicators for each element, and each indicator further described by two to four criteria.

Regarding participation, there are particular types of overt behaviors that are indicative of the skills expected of effective problem solvers. First is *action*, which describes the general level of participation of an individual, regardless of whether this action is coordinated with the efforts of others. Problem solvers differ in the level of competence with which they act in a group. While some may be passive, others become active when provided with sufficient prompts and supports, and yet others will demonstrate an ability to act independently and from their own initiative. Second is *interaction*, which refers to the capacity to respond to or coordinate actions with others, ranging from answering an inquiry, to actively initiating and arranging efforts, or prompting others to respond. Third is *task completion* skills, which refer to the motivational aspects of participation, including a sense of responsibility for the outcomes of collaborative effort.

Figure 3.1.

The structure of the framework for social skills in collaborative problem solving (Hesse et al., 2015)



Perspective-taking skills are described as the ability of a person to see a situation from the viewpoint of another person. This involves two subskill elements, namely responsiveness and audience awareness. *Responsiveness* is a set of skills used to integrate contributions of collaborators into the student's own thoughts and actions. It might involve ignoring, accepting, or adapting the contributions of others in the collaborating group. *Audience awareness* skills refer to the ability to tailor contributions to the needs of others or to make actions understood by others. It requires the ability to be aware of how to adapt behavior to make it more suitable for the collaborating partner to respond.

Social regulation skills are necessary for team members to bring a diversity of knowledge, resources and experience to a problem-solving challenge. These skills are most effective when participants know how to deal with different viewpoints and opinions. First is *metamemory*, which describes the ability to evaluate one's own knowledge, strengths, and weaknesses. Second are *transactive memory* skills, which describe a person's understanding of the knowledge, strengths and weaknesses of collaborative partners. Third are *negotiation* skills, which help participants find ways to reconcile different perspectives and opinions and to accommodate differences. This involves reaching a resolution or a compromise. Fourth are *responsibility initiative* skills, which allow individuals to take into account the different ways they can take initiative within a collaborative context. Some individuals focus mainly on their individual tasks, while others work on a shared problem representation, a strategic plan towards a solution, and regular monitoring of the group's progress.

COGNITIVE SKILLS IN COLLABORATIVE PROBLEM SOLVING

The cognitive skills for successful CPS are similar to the overt behaviors exhibited by individual problem solvers. They refer to the ways in which problem solvers manage a task at hand and the reasoning or hypothetico-deductive skills they use. Parallel to the ATC21S framework's organization for skills in the social domain, skills in the cognitive domain are grouped into two main capacities identified as *task regulation* skills and *knowledge building and learning* skills.

Task regulation skills are demonstrated by six subskill elements that are linked to an example behavioral indicator that is specified by two to four criteria. The six subskills are problem analysis, goal-setting, resource management, flexibility, data collection, and systematicity. *Problem analysis* is the ability to identify the components of a task and the information available for each of the components. *Goal-setting* is the formulation and sharing of specific sub-goals that will help to monitor collaborative problem-solving progress. *Resource management* reflects the ability to plan how collaborators can bring their resources, their knowledge, or their expertise to the problem-solving process and how they make decisions about the process of conflating data. *Flexibility* and ambiguity management skills encompass tolerance for ambiguity, breadth of focus, and communication. It might also involve the need to negotiate and to understand the perspective of other collaborative partners. *Data collection* involves exploring the task and understanding

the problem space. It requires recognition of a need for information related to the activity and understanding of how this affects and can be used by self and partner. *Systematicity* refers to the thoroughness and efficiency of the problem-solver's approach.

The learning and knowledge building skills involve many of the steps already explained in the social domain. As they progress through a CPS task, individuals can learn about content material, learn strategies and skills, learn how to deal with setbacks, or learn how to coordinate, collaborate, and negotiate with others. The ATC21S framework identifies three subskill elements within the learning and knowledge building set: relationships and patterns, contingencies and rules, and generalizing and testing of hypotheses. The *relationships and patterns* element addresses the fact that sharing and exchange are important in a collaborative setting where the partners have different amounts and types of information. There is also a need to explain these relationships to partners and to check for connections between information. *Contingencies and rules* describe the way in which the collaborators link information and, for example, communicate cause and effect. This enables them to establish simple rules which they can share in order to make progress towards a solution. Systematic observations of cause and effect enable partners to establish and discuss the potential of rules, either for the regulation of the task or the way in which they will collaborate. Rules are used to complete steps or parts of the solution. *Generalizing and testing hypotheses* demonstrate an ability to generalize by forming and testing hypotheses, using a "what if…?" approach. This is a way of describing the process and is a signal of higher order thinking and reasoning by students. It entails exploring multiple pathways to a solution. They need to be able to understand and discuss the link between action and events.

Comparison of PISA and ATC21S Frameworks

It is apparent the ATC21S and PISA frameworks for defining collaborative problem solving share a number of common features even though there are some differences in particular skills at a micro-level. Both frameworks have a collaboration dimension that involves particular social and communication competencies that end up being crossed with a problem-solving dimension that involves reasoning and other cognitive processes. The links between the components of the collaboration and problem-solving dimensions (i.e., points of intersection) have particular skills that can be measured with behaviors and acts of communication that vary in quality. Finally, both frameworks for defining CPS view assessment among its priorities.

Other Perspectives on Collaborative Problem Solving

This section discusses some other collaborative problem-solving frameworks that introduce additional factors that are potentially relevant to a NAEP CPS assessment. These frameworks may provide perspectives that improve a NAEP assessment beyond the PISA and ATC21S frameworks and may provide additional guidance in the selection of NAEP tasks. The PISA CPS 2015 framework document (OECD, 2013) also provides a review of previous research and theory on collaborative problem solving.

GROUP COMMUNICATION THEORY

There is an established line of research in the communication, cognitive, and organizational sciences that has documented communication processes for effective teamwork (Bowers, Jentsch, Salas, & Braun, 1998; Dillenbourg & Traum, 2006; Orasanu & Salas, 1993). Communication is essential for a team to collaborate, coordinate actions, provide feedback, develop strategies, and make decisions. Indeed, five to six decades ago, research on group communication covered an array of relevant topics, such as breakdowns in communication (Black, 1955), effects of competition and cooperation (Harnack, 1951), groupthink (Keltner, 1957), group developmental stages (Tuckman, 1965), and conflict within groups.

Relevant to CPS are Bales' (1953) equilibrium theory and Gouran and Hirokawa's (1996) theory on group decision making. According to equilibrium theory, teams have two specific types of needs: task needs and socio-emotional needs. The only way for teams to be effective is for there to be a balance between these needs. Bales (Bales & Strodtbeck, 1951) proposed and tested the notion that as groups solve problems, socio-emotional issues arise as a result of the tension that builds as groups work together. If teams do not find ways to release the tension by some form of positive or negative activity, then the tensions will only continue to build. If the building of tension goes unchecked, this may reduce the teamwork environment, and prohibit the progression of the team towards its goal. Gouran and Hirokawa (1996) drew upon equilibrium theory as well as other psychological theories to explain the function of communication among team members and its effect on the team's purpose. Specifically, communication can facilitate critical thinking as well as prevent a group from making errors. Individuals who view the problem from a different perspective can relay their interpretation of the situation to others so that the team can coordinate their actions, identify errors, and plan next steps. For successful problem solving, individuals within the team need to understand how to facilitate effective communication by all team members with a task focus that minimizes socio-emotional or interpersonal issues that often arise.

MACROCOGNITION IN TEAMS

Macrocognition in teams focuses on complex cognition in collaborative settings as well as the internalized versus externalized cognitive processes that occur during team problem solving (Letsky, Warner, Fiore, & Smith, 2008). It involves the knowledge work that enables the team to develop problem representations, co-construct knowledge, and generate candidate solutions for the problem at hand (Fiore, Rosen, Smith-Jentsch, Salas, Letsky, & Warner, 2010). The macrocognition in teams approach integrates three important notions. First, there are both individual and team level factors that need to be interrelated during collaboration. Second, there are both internalized and externalized cognitive functions, namely the knowledge held by individuals within the team and the artifacts created by the team in the service of problem solving and decision making. Third, it incorporates temporal characteristics to examine CPS phases and how these alter both

processes and performance. More specifically, the three notions described above are composed of five interdependent components that function in an iterative manner. These components include *individual knowledge building* (actions taken by individuals to expand their own knowledge), *team knowledge building* (actions taken by teammates to disseminate information into actionable knowledge for team members to develop applicable solutions to the problems), *internalized team knowledge* (knowledge held in the individualized minds of team members), *externalized team knowledge* (facts, relationships, and concepts explicitly agreed upon by factions of the team), and *team problem-solving outcomes* (assessments of quality of a team's problem solutions in relation to their objectives).

MODELS THAT INTEGRATE COLLABORATIVE LEARNING AND PROBLEM SOLVING

O'Neil and Chuang (2008) developed a model that simultaneously considers the collaborative components of learning and problem solving. CPS is divided into collaborative learning and problem solving. Collaborative learning in a team context can be defined, measured, and reported through six subskills (O'Neil et al., 1997): *adaptability* (recognizing problems and responding appropriately), *coordination* (organizing group activities to complete a task on time), *decision making* (using available information to make decisions), *interpersonal* (interacting cooperatively with other group members), *leadership* (providing direction for the group), and *communication* (clear and accurate exchange of information). In turn, problem solving has three components. *Content understanding* is the domain knowledge required to solve a problem. *Problem-solving strategies* can be domain dependent or domain independent. *Self-regulation* involves motivation and metacognition, each of which are further subdivided. Motivation consists of effort and self-efficacy, whereas metacognition consists of self-checking and planning.

ADDITIONAL FRAMEWORKS AND GAPS

While many approaches to CPS either implicitly or explicitly differentiate between the cognitive factors of problem solving and the social factors of problem solving, there exists a rich body of literature from which to draw that may enable a finer level of granularity for defining CPS processes. In particular, in the organizational sciences, team training research provided an important distinction between what was defined as taskwork and teamwork (Bowers, Jentsch, & Salas, 2000; Cannon-Bowers, Tannenbaum, Salas, & Volpe, 1995). Taskwork is a label for the activities in which one engages that are pertinent to achieving the goals and objectives for which the team is formed (e.g., running a procedure for data collection, completing a particular statistical analysis). Teamwork describes the activities involved in interacting with members of one's team and that are necessary for success (see Fiore, Carter, & Asencio, 2015). Regarding teamwork, there is knowledge associated with teammates with whom one is collaborating. This includes understanding the roles, responsibilities, and agents' capabilities in executing these (see column 3 in Table 3.1). Second, there are the skills supporting interaction with teammates during problem solving, such as communicating effectively about a project and managing conflict. Third, there are attitudes about teammates based

upon interactions, including trust in teammates and a sense of cohesion with them. These skills are identified in the Bowers' model of CPS (Fiore, Carter, & Asencio, 2015).

Adequate conceptualization of a problem is instrumental to the problem-solving process (Chi, Glaser, & Rees, 1982; Newell & Simon, 1972), particularly in teams (Fiore & Schooler, 2004). Problem conceptualization, or the development of a problem space, is the stage where the problem solver encodes the salient aspects of the problem at hand such as the characteristics of the environment, the goals of agents relevant to the problem, as well as rules for operating (Newell & Simon, 1972). Originating in human information processing theories, problem space theory has also been developed in the context of teams (e.g., Hinsz, Tindale, & Vollrath, 1997; Orasanu, 1994). A key element of success during the problem conceptualization stage, is shared cognition. That is, some level of overlap between team members' understanding of the essential problem characteristics is a mandatory factor that contributes to a team's ability to build representations that are effective in the generation of quality problem solutions (Fiore & Schooler, 2004). As such, problem conceptualization is an important area of research for developing our understanding of CPS assessment.

All of the CPS frameworks acknowledge that proficiency in both individual and collaborative problem solving is robustly predicted by the individuals' prior knowledge of the subject matter, called *domain knowledge* (see also the NRC, 2012). General problem-solving skills do uniquely predict individual problem-solving performance over and above domain knowledge and other cognitive abilities, such as numeracy, literacy, general intelligence (Greiff et al., 2013), but a sufficient base of accessible, well-integrated knowledge may be necessary for sophisticated problem-solving methods to be applied. Someone who is a very capable problem solver in one area may be unable to transfer their problem-solving skills and strategies to another domain where they lack expertise. These conclusions on individual problem solving may also apply to the collaborative dimension of CPS. The ability to communicate information and assign roles to team members would undoubtedly be influenced by domain knowledge. The selection of problem contexts in PISA CPS 2015 directly considered the role of domain knowledge; the test takers were expected to have sufficient subject matter knowledge to solve the problems.

Team members are more effective if they reflect together upon the problem-solving activities before, during, or after the completion of tasks and subtasks (West, 2000; Schippers, Den Hartog, Koopman, & Wienk, 2003). The concept of *team reflectivity* refers to these processes, which are captured by the monitoring and reflection components in the PISA framework (row 4 in Table 3.1). Reflection can be encouraged or required in the formal structure of the task, as in the case of evaluations, debriefings, and after-action reviews that are common in industry and the military. Digital technologies can also encourage reflection by requesting evaluations, feedback, and recommendations by team members at different phases of problem solving. These formal and technological artifacts encourage team reflectivity in ways that are not routinely exercised when teams solve problems following their own proclivities.

In real-world problem solving, teams must often adapt when dealing with changes to the problem or when addressing interpersonal issues that cause friction (Gurtner, Tschan, Semmer, & Nägele, 2007). Data suggests that effective teams are able to draw upon a rich repertoire of strategies that support adaptive performance (Entin & Serfaty, 1999). Overall, teams adapt by gathering information, making judgments, and selecting an appropriate response (e.g., Cannon-Bowers, Tannenbaum, Salas, & Volpe, 1995; Mosier & Fischer, 2010). Adaptive collaborative problem solving will likely rely on many of these team processes (Mosier & Fischer, 2010; Orasanu, 2005; Orasanu & Salas, 1993). This suggests a need for adaptive testing in general, but, if possible, testing adaptive to particular subsets of skills that can differently manifest in CPS.

Another key challenge for CPS is the need to manage the group dynamics that emerge when team members come from varied cultural or disciplinary backgrounds, have different commitments to the team (e.g., unequal commitments of a time on a project), have different personalities, or have different degrees of authority. Multiple forms of expertise from the task standpoint must also be managed. Conflicts among team members produce complex group dynamics that cut across cognitive, behavioral, and affective levels. This is apparent in the early stages of problem solving, when problem identification takes place. It is also apparent in the middle stages when there are different perspectives on appropriate strategies, assignments of tasks, adoption of roles, and re-planning as obstacles occur. When OECD commissioned the PISA CPS framework, the expert group was instructed to focus on the cognitive and social aspects of CPS rather than the personalities and emotions of team members. It is worthwhile to revisit how personality and emotions contribute to team performance in future CPS assessments.

Conclusion

In this chapter we have described some of the primary frameworks for CPS, as well as discussed additional approaches that can add to assessment. Defining collaborative problem solving will provide a number of important insights for multiple stakeholders. At the most foundational level, this chapter shows how it is possible to leverage theory and concepts from the cognitive and learning sciences and meld them more broadly with knowledge from the social and organizational sciences on teams. This work has implications cutting across industry, government, and the science and technology sector. Team science is not confined to a particular field as it is increasingly practiced within and across a variety of disciplines cutting across the physical, social, life/health and computational sciences (e.g., Fiore, 2008; Lazer et al., 2009; Stokols et al., 2008). As such, as we have seen happening with organizations more generally, the process of collaboration is becoming the norm rather than the exception. The definitions and concepts discussed in this chapter are provided to help stakeholders understand the contexts, structures, and processes of collaborative problem solving through a richer understanding of team and organizational processes. This is a necessary precursor for developing a model for future education and training in collaborative problem solving.

4 Assessment Design

Introduction

There are many, diverse examples of collaborative problem-solving tasks in use in schools in the United States and elsewhere, ranging from informal classroom activities to large-scale formal evaluations of collaboration through expensive online training systems (Griffin & Care, 2015). There is also extensive research literature on the factors that affect the success of collaborative learning and collaborative performance (von Davier, Zhou, & Kyllonen, 2017) and discussions of design principles (e.g., O'Neil, Chuang, & Chung, 2004; OECD, 2013; Kyllonen, Zhu, & von Davier, 2017). This literature includes evaluations of factors such as the content domain, students' familiarity with the material, collaboration incentives, and team composition on individual student and team outcomes. The purpose of this chapter is to identify the factors that should be considered in designing tasks for the assessment of collaboration, so that they might be manipulated or controlled in designing a collaborative assessment for NAEP.

This chapter has five sections. The first section discusses the nature of the collaborative task, including the content domain, whether the problem presented to the team to be solved is ill-defined (i.e., without clear goals or expected solutions) or well-defined (i.e., with specific goals and expected solutions), familiarity of the content to participants, and incentives for collaboration. The second section discusses the issue of human-to-human vs. human-to-agent collaborations, and the number of team members in a collaboration. The third section discusses team composition issues, such as whether team members are similar or dissimilar in ability, gender, and other factors, and whether or not they assume or are assigned different roles. The fourth section discusses evaluation issues, such as whether individual or team outcomes are the primary reporting variable, and whether process variables are considered in evaluation of team effectiveness. A range of potential process measures is proposed, and variables that might moderate the strength of the relationship between background variables (person, situation, and process) and outcomes are suggested. The final section includes some conclusions and considers several tasks with respect to design features.

As with all NAEP assessments the nature of the task will affect which students, schools, and states do relatively well on them and which do poorly. However, unlike the case with mathematics, English/language arts, and science, there is no professional-organization-backed or national set of K-12 standards for collaborative skill development, at least that we are aware of. We therefore cannot rely on an existing, agreed upon set of standards to provide content and design specifications for a collaborative task. Therefore, the goal for this chapter is to identify the set of factors that have been shown to make a difference in collaboration, which will enable us to design a collaborative task in consideration of those factors. In evidence-centered design (ECD) terminology (Mislevy, Steinberg, & Almond, 2003), the process of identifying these key factors is referred to as a *domain analysis* (see Table 4.1, for a listing of ECD phases and related terminology). Having conducted the domain analysis, we then can be in a position to propose a set of collaborative-task *design patterns* or *test specifications* (see Table 4.1), and have a basis for evaluating their merits with respect to the kinds of claims about collaborative proficiency we might want to make.

Table 4.1.

Evidence-Centered Design (ECD) phases and related terminology (Mislevy, Steinberg, & Almond, 2003)

ECD Phase	Related terminology
Domain analysis	task analysis, job analysis, identifying content standards
Domain modeling/Design patterns tool	test specifications, KSAs
Conceptual analysis framework	measurement model, scoring model, delivery platform
Assessment implementation	item writing, test assembly, scoring, task modeling
Assessment delivery	field trial, pilot test, operational data collection, score reporting
Design Patterns Tool	**Related terminology**
KSAs (primary, secondary)	knowledge, skills, abilities, competencies
Benchmarks	standards, cut scores, proficiency levels
Rubrics	coding schemes
Characteristic & variable features	item features, radicals/incidentals (automatic item generation)

Nature of the Collaborative Task

Several major dimensions of collaborative learning and problem solving have been suggested. One is the type of interaction between team members: individualistic, cooperative, or competitive. Individualistic tasks are those on which students work together but their inputs are not combined; cooperative tasks are those that require the combination of students' work to complete the task; and competitive tasks are those that encourage students to outperform others (Arterberry, Cain, & Chopko, 2007). Most of the literature on collaborative problem solving is focused on *cooperative tasks*. This is likely the most appropriate kind of task for a NAEP collaborative assessment.

Another important aspect of the nature of the task is its content. Collaboration could be assessed on any number of cognitive tasks, including the kinds that are already administered in NAEP assessments (e.g., mathematics, reading, science, technology and engineering literacy, art). However, tasks could also reflect problem-solving experiences outside curricular areas, or outside the areas

for which there are NAEP assessments, such as leisure time activities, social and interpersonal activities, intrapersonal behavior (personal time management, anxiety reduction), and so on. McGrath (1984) provides examples of a wide variety of collaborative tasks through his "group task circumplex" which arrays tasks on a conflict vs. cooperation dimension and an orthogonal choose (conceptual) vs. execute (behavioral) dimension. Task types and example tasks are listed in Table 4.2.

Table 4.2.

Varieties of collaborative tasks based on McGrath's (1984) group task circumplex

Quadrant/Task Type	Subtask Type	Examples
I. GENERATE (ideas or plans)		
1. Planning Tasks[a,d]	Generating plans	Agenda setting
2. Creativity Tasks[a,c]	Generating ideas	Brainstorming
II. CHOOSE (an answer to a problem)		
3. Intellective tasks[a,c]	Solving problems with correct answers	Logic problems
4. Decision-making tasks[b,c]	Deciding issues with no right answer	Appropriate sanction for rule violator
III. NEGOTIATE (a solution to a conflict)		
5. Cognitive conflict tasks[b,c]	Resolving conflicts of viewpoint	Prisoner's dilemma
6. Mixed-motive tasks[b,d]	Resolving conflicts of interest	Negotiation tasks
IV. EXECUTE (behavior)		
7. Contests/battles/competitive tasks[b,d]	Resolving conflicts of power	Sports
8. Performances/psychomotor tasks[a,d]	Executing performance tasks	Psychomotor tasks

NOTES: [a]Cooperation Tasks; [b]Conflict Tasks; [c]Conceptual Tasks; [d]Behavioral Tasks
Examples from Straus (1999, tasks 1 to 4) and McGrath (1984, tasks 5 to 8).

Problem-solving tasks may present well-defined or ill-defined problems. Well-defined problems are ones that comprise clear goals, unambiguous procedures, and a correct response (see Task Type 3, Table 4.3). For such well-defined problems, a common finding has been that students tend to exchange only information and explanations. Ill-defined problems are ones that do not have clear goals or a necessary correct answer. For such ill-defined problems, particularly ones that are more conceptual and contain open-ended responses (see Task Types 2 and 4, Table 4.3), students tend to focus more on an exchange of ideas and strategies and have a higher level of collaboration overall (Cohen, 1994; Shachar & Sharan, 1994). This suggests that to elicit more collaboration, it may be useful to consider a more ill-defined, open-ended task approach to investigating collaborative problem solving in order to focus on idea sharing.

Motivation is an important determinant of success on cognitive tasks (e.g., Liu, Bridgeman, & Adler, 2012). To decrease the possibly confounding effects of motivation on task performance, in which some students are motivated and others are not (in which case the assessment would be measuring students' motivation rather than their collaborative skill, per se), it may be useful to control for motivation. One way to do this is by making the task more engaging for all participants in order to elicit maximum effort from team members. There is not a definitive answer in the literature on how

to maximize the engagingness of a problem-solving task as a way to increase students' motivation, but there are some suggestions about factors that may tend to increase engagingness. Task novelty to invoke curiosity and interest (Baranes, Oudeyer, & Gottlieb, 2014; Hidi & Harackiewicz, 2000; Lowenstein, 1994, Malone, 1981), appropriate challenge or difficulty level (Baranes et al., 2014; Malone, 1981), feeling of being in control of success and failure (Eccles & Wigfield, 2002), incentives such as score keeping or "quantifiable outcomes" (Salen & Zimmerman, 2004, p. 80), and provision of feedback (Hattie & Timperley, 2007), or pay (Jackson, 2009) have all been invoked as motivators. While these kinds of factors are likely to affect motivation in any kind of task, including standard NAEP tasks, it might be that due to the novelty of group problem solving, motivations might play an especially important role in governing individual and team performance.

Human-to-Human vs. Human-to-Agent Collaborations and Team Size

The issue of human-to-human vs. human-to-agent collaboration is discussed elsewhere in this report. We include it here only to emphasize that this issue has design implications, perhaps the most important of which is the issue of group composition. An agent's personality, gender, ability level, and other background factors are in principle possible to express and to manipulate, but in reality, given the current state of artificial intelligence, such an expression is difficult to achieve in a believable way. Developments in this area are progressing, but the ability to reliably present an agent that responds similarly to the way a human does remains in the future. The state of the art is that a moderate percentage of people can be fooled into thinking a computer agent is human (one third, in a recent University of Reading, 2014, competition), but the ability to produce this kind of agent, one capable of passing the "Turing Test," is at the edge of computational sophistication. Nor have there been any demonstrations that two such agents could be differentiated by background factors (e.g., having agents that could both be believable as humans and be reliably differentiated as males or females). This suggests that at this point true collaboration may still be a human-to-human activity, although some literature suggests that well-constructed human-to-agent collaborative tasks capture many features of human-to-human collaboration, and the two task types have similarities (e.g., in motivating students, in eliciting communication, in overall student performance) as well as differences (e.g., humans disagree more with agents than with humans) (Rosen, 2015).

There are several current examples of human-to-human collaborative assessment systems being used in schools or for research purposes. These include the ATC21S collaborative task (Griffin & Care, 2015), ETS's Tetralogue system (Hao, Liu, von Davier, & Kyllonen, in press), and Rosen's (2015) system which was designed to compare human-to-human with human-to-agent approaches. ETS's Tetralogue system (Hao et al., in press) also might be considered a hybrid because the system hosts human-to-human interactions, via a chat window, but also includes a computer-based deterministic facilitator in the form of a teacher agent and fellow student agent.

If a human-to-human collaborative task is designed, then an issue is how many students should be assigned to a collaboration task. Students could be assigned to dyads, triads, tetrads, pentads, or larger groups. One advantage of dyads is that they are efficient per person in providing more group outcome data points (N/2) than other arrangements (i.e., N/3, N/4). (That is, if N = 24 students in a school participate in an assessment, dyads will provide N/2 = 12 team outcomes, triads N/3 = 8, and tetrads, N/4 = 6.) However, a triad might have qualitatively different dynamics from a dyad. For example, several different cliques are possible with triads (AB, AC, BC, ABC) but only one is possible with dyads (AB). Clique dynamics could be an important part of the collaborative problem-solving experience (e.g., Cooke, Gorman, & Winner, 2007). Also, social loafing—the phenomenon of a team member being less motivated, exerting less effort compared to individual performance, and relying on others in the team to contribute to the team outcome—would seem to be more possible as team size increases. There is some literature suggesting that this is true (Karau & Williams, 1993, see footnote 9). It may be useful to conduct research comparing dyads and triads.

Team Composition

CPS happens in a team setting. Whether the goal of its assessment in the large-scale context is to measure the ability of an individual or a team, the performance of an individual inevitably depends on the composition of the team in which the individual is participating, and this is why team composition is an important factor to consider in developing a CPS assessment.

Team composition is determined by several characteristics that we discuss in this section: team members' (a) ability and knowledge, (b) roles, and (c) background characteristics, including gender, cultural background, and socioeconomic status (NCES, 2012).

All of these team characteristics—abilities, roles, background—may affect collaborative problem solving at both the individual and team level, and therefore it is important to consider them. For instance, a student at a given ability level might perform differently depending on whether he or she were placed in a low- or high-ability group. And two teams whose team members have the same skills and abilities on average might perform differently depending on the mixture of skills on the team. In the real world, individuals need to be able to work with various groups (OECD, 2013). For this reason, and for the reason of sampling different compositions, it may be useful to evaluate designs that mix team compositions randomly or systematically, either by having an individual join more than one team for an assessment, or through some other means, such as composing teams according to an experimental design.

DISTRIBUTIONS OF ABILITY AND KNOWLEDGE

In any team activity, each member comes with certain abilities, skills, and knowledge. This distribution may be homogeneous, heterogeneous-hierarchical, or heterogeneous-nonhierarchical.

Homogeneous teams are formed by members with similar background, skills, and knowledge, (Webb, 1991). For example, for the purposes of evaluating performance on a mathematics problem solving exercise, teams homogeneous in their mathematics ability could be assembled based on a mathematics pretest, the results of a previous standardized test, course-taking history, grades within courses, or by teacher ratings. In a NAEP context, students randomly identified for NAEP participation could be assembled into teams based on matching such scores, history, or ratings.

From an assessment standpoint, the advantage of a homogeneous team is that the team's ability level and the ability level of its members are approximately the same (e.g., high ability) and so a single variable can represent the team's ability level. This simplifies interpretations regarding the relationship between ability and performance. (With heterogeneous teams, it may take several variables to capture the team's ability, such as individual team members' abilities, the average ability, the highest ability; see Woolley, Chabris, Pentland, Hashmi, & Malone, 2010).

A potential disadvantage to a homogeneous team is that it may encourage social loafing (Karau & Williams, 1993), in which a member leaves the work to other team members. Homogeneous teaming (with teams of a size greater than two) might encourage such behavior because a member might see participating as unnecessary, since the other members know more or less the same as he or she does. That is, if three members have similar skills and knowledge, the collaboration may be perceived as unnecessary by one of the members. Social loafing could occur in heterogeneous teams, too (e.g., a low-ability student might want to remain passive while two high-ability team members solve the problem). But if it turned out that a common cause for social loafing was perception of knowledge overlap among team members, then a task design modification known as the *jigsaw* (American Psychological Association, 2003), in which team members are each supplied with unique information relevant to problem solution, could be a remedy.

Heterogeneous teams are composed of members with varying levels of knowledge, skills, and ability with respect to the content of the problem-solving task. This is a natural team in a typical school setting in the sense that students randomly assembled together in a team will tend to vary in their problem-solving ability; that is, teams will be a mixture of strong to weak students. On such heterogeneous teams, the stronger students—those with more knowledge or ability—might play a leadership role as a result of their recognizing that they have the knowledge to contribute to the problem-solving task, and may also recognize that they are in a position to mentor less knowledgeable members. The weaker students benefit from the peer mentoring (Webb, 1995). Webb (1998) showed that heterogeneous groups tend to be more successful in problem-solving

activities than homogeneous groups, and that heterogeneous groups are especially advantageous for low-ability students when mixed with high-ability students. From an assessment perspective, the point is that the group dynamics in homogeneous versus heterogeneous teams might be different, and therefore different outcomes might be observed.

Cross-functional teams (Parker, 1994/2003) are a type of team widely observed in a workforce context in which members with different expertise are assembled in pursuit of a common goal. These teams are commonplace and often effective in a workforce context because project goals often require different kinds of expertise for their successful completion. For example, a team might consist of members from the research department, the IT group, the marketing division, the legal office, and a business owner. Such teams can be effective because each member brings a different type of expertise and can offer a unique contribution to the solution, and different aspects of the problem and possible solutions are discussed collaboratively. Such an arrangement does not typically apply to a school setting; however, the jigsaw method can simulate team members bringing different necessary expertise. And it may be that non-academic subject matter (such as fantasy or generating solutions to absurd problems) that would enable each student to draw on idiosyncratic personal experiences might have a similar effect.

ROLES

In CPS tasks, it is possible to have emergent or prestructured groups. In *prestructured* CPS groups, roles are assigned in advance, and each member enters the CPS task with an already determined role. In *emergent* CPS teams, the roles are not pre-assigned, but teams are formed spontaneously and the roles evolve over the course of the CPS activity.

Depending on their content, CPS tasks can encourage or discourage leadership and role-taking behavior in general. For example, in jigsaw problems, where each member has a piece of information necessary for the solution, the participation of each member is encouraged so that it is more likely that all members will be contributors, and a member might even take the role of monitoring the group's progress. In a team with heterogeneous abilities, it may be more likely that leadership versus follower roles will develop according to abilities.

Assigning roles in advance could encourage students to show specific behaviors and could thereby provide insight into a wider range of students' CPS skills. For example, if a student were assigned the leader role in one task and a follower role in a second, the assessment could provide information about the student's behavior in different situations. However, if the roles in each task were emergent, it is possible that the student would tend to take on the same role, and the assessment would provide an incomplete picture of the capabilities of a student in contributing to the solution of a CPS task.

MEMBER CHARACTERISTICS

Past research has shown that the performance of the group depends on various characteristics of the group members. These characteristics consist of gender or socioeconomic status, among others. With regard to gender, although Webb (1991) found that boys tended to ignore girls in mixed-sex CPS groups, Woolley et al., (2010) found that the more (higher percentage of) females on a collaborative problem-solving team, the better the team performance was. These results indicate that the distribution of gender may play a role in the CPS performance of students and needs to be considered when designing an assessment.

The symmetry of socioeconomic status can also be an important aspect of CPS groups. Collaborative work between peers will likely be different than CPS between, for instance, teacher and student (e.g., Lai, 2011) and can possibly trigger different processes depending on the culture. Research indicates that low status students manifest less participation and have less impact on the work process than higher status students (e.g., Cohen, 1982, 1994; Cohen, Lotan, & Catanzarite, 1990). These and other group member characteristics, which Webb (1995) calls *status characteristics* (e.g., race, popularity, perceived intelligence), can all play an important role in the constellation of the team, team performance, and the team processes that can be observed.

Evaluations

This section addresses the issue of what is to be evaluated in a collaborative problem-solving assessment. There are several methods employed when evaluating collaborative group work (Schmitz & Winskel, 2008)—individual outcomes, group outcomes, and individual or team process measures. From the adult literature on group problem solving, there is the suggestion of major dimensions of collaboration, a "Big 5" of teamwork: team leadership, mutual performance monitoring, backup behavior, adaptability/flexibility, and team orientation (Salas, Sims, & Burke, 2005).

INDIVIDUAL OUTCOMES

An individual outcomes approach to evaluating collaborative problem solving focuses on the results from the collaborative experience. Individual students are given a pretest (e.g., as part of the collaborative task experience in the CPS platform, but performed alone without the collaborators), then engage in a collaborative activity. They are then given an individual post-test. An individual outcomes score is computed by subtracting or regressing out the pretest score. The pretest is used to ensure that the outcomes score reflects the collaboration per se and not prior knowledge or skill. This method is relatively economical (it requires only a pretest and post-test item or scale) and enables large-group testing, and could therefore fit into a NAEP data collection. The method can produce a collaboration score even while ignoring the process data generated by the interactions that occur between students during the collaborative task (Griffin, 2017).

TEAM OUTCOMES

The team outcomes method assesses the problem-solving team, not the individual, as the unit of analysis. Such an approach is novel, particularly in the context of a large-scale cognitive assessment such as NAEP, but it may be the one that best reflects team performance per se. If the goal is to assess team performance, rather than individual performance, and to report on or make claims concerning team performance in schools, districts, states, and in the nation, then a team score may be the best indicator and basis for such claims. The industrial-organizational psychology literature provides suggestions for how team problem solving can be conceptualized and assessed. A useful distinction is between teamwork knowledge, taskwork knowledge, and team situational awareness, all of which predict task outcomes and process behaviors (Cooke, Salas, Kiekel, & Bell, 2004). Students' team mental models (Cannon-Bowers, Salas, and Converse, 1990), which guide information sharing and assist transactive memory and knowledge of team learning and team consensus may be useful to measure (Mohammed & Dumville, 2001). It may also be useful to adopt a teamwork competencies taxonomy in which knowledge, skills, and attitudes are specific or generic to a task and specific or generic to a team (Cannon-Bowers & Salas, 1997).

PROCESS MEASURES

A process measure approach focuses on within-group interactions, mainly the conversations or chats between team members, but also potentially other information, such as diagrams, gestures, and other artifacts (these may be especially available if there is a relaxed pace to the problem solving; see Mathieu et al., 2000). Process measures would include items such as the number of statements made during problem solving, the amount of turn taking (transition from one person to another person talking, texting, or chatting), and the levels of participation of each member (e.g., the amount of pertinent discussion contributed) (von Davier & Halpin, 2013; Liu, Hao, von Davier, Kyllonen, & Zapata-Rivera, 2015; Woolley, Chabris, Pentland, Hashmi & Malone, 2010). If statement content can be analyzed, then additional process measures could include the number of personal acknowledgement statements, the number of statements made pertaining to planning and goal setting, and the amount of comprehension monitoring statements, such as questioning and answering, elaborations, explanations, summaries, or comprehension monitoring activities, such as making a diagram. A content analysis of process interactions might try to characterize activities, such as recognizing contradictions in the problem-solving process and resolving them, and learning and enacting effective problem-solving strategies. Team members could also be assessed on their knowledge of teamwork itself, particularly how it changes over the course of the problem solving (Cooke, Salas, Cannon-Bowers, & Stout, 2000).

A process measure approach, compared to an outcomes measure approach, is more labor intensive to score, and more preparation is needed to determine how best to code interactions compared to an outcomes method (although progress is being made; see Adams, Vista, Scoular, Awwal, Griffin,

& Care, 2015; Hao, Liu, von Davier, Kyllonen, & Kitchen, 2016). Computer administration makes possible the automatic recording, and coding of process information. However, considerable research is necessary to determine how to do this with all the potential process measures that could be evaluated. A summary of much of what we know about team performance (Salas, Cooke, & Rosen, 2008) concluded that we need better measurement, particularly practical, unobtrusive, real time measures.

MODERATORS

There are a number of potential moderators that should be considered in the evaluations, such as demographic characteristics and aspects of the content of the assessment. Demographic characteristics include the major NAEP reporting variables: age and grade, gender (male-female), race/ethnicity, socioeconomic status, English language learning status, and disability status, along with teacher and school variables. The content of the problem solving might also be considered as a potential moderator. This would consider the degree to which the material engaged the students and the curricular content of the material (e.g., mathematics, English/language arts) or non-academic content (e.g., leisure activities, everyday activities).

Conclusion

There are many design decisions that could potentially affect the kind of collaborative problem-solving task that could be administered in the NAEP program. These design decisions are important because they would determine the types of inferences about student collaborative skills that would be made on the basis of individual or team performance. The key design decisions fall into the categories of (a) whether the task is designed for human-to-agent or human-to-human collaborations; (b) the nature of the task, such as its content domain and whether it is an ill- or well-defined problem; (c) team composition, such as whether teams are homogeneous or heterogeneous with respect to age, ability, gender, and other factors, and whether roles are assigned or spontaneous; and (d) how collaborative problem-solving ability will be evaluated—as individual outcomes, team outcomes, process measures, or some combination of the three.

NAEP may consider designing a CPS task with the following characteristics: a task for human-to-human collaboration that is (a) between two humans, so that the number of independent groups is N/2 (rather than N/3 or N/5); (b) cooperative (not competitive or independent); (c) ill-defined in order to encourage collaboration; and (d) engaging so it increases students' motivation to perform well. NAEP might also consider that a variety of team composition types (e.g., heterogeneous, assigned roles) be manipulated to provide collaborative experiences that reflect real world collaborations, and that may include predefined roles (e.g., leader, follower) to encourage collaboration. In this design, individual and team outcomes could be measured, and process variables could also be measured. Some process variables (e.g., number of individual contributions, turn taking) can be

generated even without a content analysis, but additional process measures (e.g., goal setting, comprehension monitoring, and time off task) can be produced by examining content. Team membership characteristics—including gender, race/ethnicity, and socioeconomic status—should also be considered in analysis and manipulated in assembling teams. Research programs are likely to be needed to explore these options.

The choices made in the process of assessment design regarding these various task aspects will trigger different CPS processes. For this reason, test developers need to consider how to structure the assessment to cover all the facets named in the framework of CPS and to ensure the validity of the assessment. The main goal of developing a collaborative assessment task for NAEP is to produce a task that serves the role of eliciting evidence for student collaborative skill that is valued in education and in the workforce. The nature of this skill is an important topic for further discussion.

5 Assessment Development

Chapter 3 described two conceptual frameworks in depth (PISA 2015 and ATC21S). Chapter 4 described different considerations and approaches to the assessment design. This chapter describes the assessment development approaches taken by these two assessments.

PISA 2015 CPS Task Development

TASK CHARACTERISTICS

In human-to-agent assessment of CPS, individual students interact with one or more agents during the process of solving a problem. Each test taker participates in three to four different problem scenarios that are divided into two 30-minute sessions on a computer. The scenarios have different team compositions and partners with different characteristics (e.g., cooperative, negligent, error-prone) so that the test taker is assessed systematically in different conditions that are aligned with the 12 cells in the CPS proficiency matrix. The entire PISA assessment had six scenarios with one to three agents per scenario, three 30-minute units, and one to three scenarios per unit.

The scenarios had a number of common characteristics that the developers followed in the computer based assessment. Examples of the scenarios are provided in the CPS Framework and on the OECD website. For example, in one problem, the main goal is to find the best conditions for fish to live in an aquarium. Two team members (the test taker and a computer-agent peer) work together to explore the impact of six factors on the conditions in the aquarium. The test taker controls three factors (water type, scenery, and lighting) and the agent controls the remaining three factors (food type, number of fish, and temperature). The student and agent need to collaborate in order to solve the problem within particular time constraints and cannot solve the problem alone given that each team member controls different factors.

The computer screen has two major regions: one that displays a problem work area and the other that displays a chat interaction area. All of the problems in the assessment have both physical actions that the test taker performs in the action area and verbal actions in the chat area. Moreover, these two forms of test taker input are simplified so the computer can automatically score the test

taker's behavior. The physical actions are clicks on particular locations in the action area (such as a location on a map or the selection of an object); the verbal actions are selections from three to four alternative messages in a chat menu. One important constraint is that there are a small number of options (two to six) that the test taker has available for each physical action and each chat message; this is analogous to the standard multiple-choice test formats in assessment methodology. Each scenario has 10 to 30 common critical points during the problem-solving scenario when the computer collects the test taker's actions or chat decisions. Each of these critical points is aligned with one of the 12 cells in the CPS proficiency matrix.

The task developer needs to control and anticipate all of the events and stimuli that the test takers are exposed to and that the test takers can trigger by their own actions. The sequence of the computer-agent's actions, messages, or the change of states in the task are programmed in advance. Each test taker is confronted with the same problem and with a common set of critical points for assessment, thereby offering standardized conditions that are necessary in large-scale assessments. The test taker's selections among options at the common critical points contribute to the scores, as defined by the CPS proficiency matrix. Additional details on managing conversations with agents to meet these constraints can be found in other reports (Graesser, Foltz et al., under review; Graesser, Forsyth, & Foltz, in press).

Important Features of the PISA Tasks

The PISA framework identified a large number of scenario features to be considered in the set of problems to develop. Some features ended up being held constant, whereas others varied across problems. For example, it was decided that while the roles of the human and agents may differ, the status of all participants must be equal. This was decided so that countries are not penalized if their culture does not allow lower status team members to ask questions of higher status team members. Another very important constant feature was interdependency. As illustrated in the aquarium example, each collaborator has control over some but not all variables in the task (i.e., three for the test taker and three for the agent) so the test taker cannot solve the problem alone and members of the team are forced to work together. Whereas status and interdependency are features that were held constant in PISA, others were varied. For example, the number of computer agents varied from one to three in PISA because there was a question as to whether additional CPS complexity would lead to problems with more agents. Another feature that varied was the difficulty of the problem to be solved. An important step in the development process is to decide which features of the problems should be held constant and which should vary.

The aquarium problem also illustrates a frequent phenomenon in CPS tasks called the *hidden profile* problem. The problem starts out without the test taker knowing about the control panel of the computer-agent. Thus, the team members start out not knowing what each other knows and can perceive. They need to have a conversation in order to establish shared knowledge (i.e., common ground) (Clark, 1996; Dillenbourg & Traum, 2006), one of the main CPS competencies. The chat

facility presents the chat options for the test taker to reconstruct the knowledge and perspectives of other team members. An important step in the development process is to select problem scenarios with characteristics that allow physical actions and chat messages that tap many of the cells in the proficiency CPS matrix. However, it is not necessary for a single problem scenario to cover all of the cells in the matrix. A particular cell may be covered by some scenarios but not others.

The PISA framework identified some context dimensions that could be considered in selecting the problems. The dimensions of the problem scenario specify the type of problem-solving task, such as hidden profile problems, jigsaw problems, consensus building, collaborative design, negotiations, and so on. Hidden profile, jigsaw, and consensus were the categories of choice in PISA 2015. Problem developers attempted to minimize reliance on domain knowledge (e.g., mathematics, science, reading), but some rudimentary skills in each of these were expected. The hope was that CPS would be correlated with scores on reading, mathematics, and science, but would explain additional variance that was unique to CPS and, thus, to the test takers' proficiency in solving problems collaboratively.

The agents in the different problem scenarios exhibited varying abilities and difficulties to the test taker. That is, the test takers worked with agents that were helpful, incorrect, social loafers, defiant, impetuous, and so on. This variability in agent design made it possible to collect a wide range of indicators of how test takers handle typical team challenges.

The Choice of the Communication Medium

As in the human-to-human (HH) assessment of CPS, the medium of communication in human-to-agent (HA) can include open-ended verbal messages, predefined chat options, and the use of audio and video media. PISA had to collect data from students in dozens of languages, cultures, and countries so there were substantial logistical constraints. For example, it was impossible technologically to automatically score open-ended verbal messages generated by the test taker in either text, audio, or audio-visual media. With respect to the conversational agents, it was undesirable to have them communicate via spoken messages and a visual persona because those depictions vary substantially among cultures and would run the risk of introducing unwanted biases in the assessment. PISA used minimalist agents that consisted of predefined chat and email messages without any speech, facial/body animation, or distinctive visual persona.

DEVELOPMENT PROCESS

Stakeholders in the Process of Developing PISA Items

The process of developing PISA items was a complex effort orchestrated by OECD, with the involvement of many countries and multiple corporations. For example, the first phase of writing the PISA framework document was led by Pearson Education, whereas the item development phase was led by Educational Testing Service. The PISA Expert Group participated in both the framework and

item development phases. During the 5-year process, each country in OECD had the opportunity to propose CPS problems for consideration, to comment on the framework document, to vote on approving the framework document, to comment on the developed items, and to approve each CPS problem. Each problem was prepared in the language(s) of the citizens for each country; accuracy of the language translations was triangulated by comparisons to both an English and a French version. The Expert Group periodically met with the item developers, gave recommendations on item development, eventually approved the items, and helped construct a tentative CPS proficiency scale based on field testing data that was collected in 2014. ETS led the collection of data, which was completed in 2015, and began the statistical analyses in 2016. An initial specification and description of the CPS proficiency scale was prepared in May 2016 in a meeting between ETS and the Expert Group who examined data for all of the items in the six PISA problem scenarios.

Iterative Development of Problem Scenarios with Human-Agent CPS

Stakeholders of the PISA assessment identified the types of problems and context dimensions in the assessment. There was the opportunity for research teams in different countries to submit candidate problems and items for consideration. Eventually, a set of prototype problems were explored and elaborated upon by the item writers at ETS. After that point, there were several steps of iterative prototyping and testing in the development of CPS items. Some of these methods were adopted by ETS, but this chapter identifies a broader set of steps that were recommended for NAEP in CPS assessment with computer agents.

1. ***Collecting verbal protocols on human-to-human interaction during CPS on problem scenarios***. This may include computer-mediated communication, face-to-face interactions, or think aloud protocols during problem solving. These uncover strategies, collaboration approaches, and distributions of actions and speech acts.
2. ***Identifying critical points in the problem-solving process.*** These are points in the problem-solving process that are aligned with cells in the CPS proficiency matrix. All test takers would receive these critical points and be assessed.
3. ***Developing multiple-choice items at each critical point***. There are a limited number of options (two to six) for the test taker to select from at each critical point, which is defined as an item. The options can either be physical actions or verbal actions in a chat facility.
4. ***Analyzing the distribution of assessment opportunities in the scoring rubric***. For PISA, the scoring rubric consists of the 12 skills in the CPS proficiency matrix. There needs to be an adequate number of observations (items) for each cell. There needs to be some index of the reliability of experts when they assign particular multiple-choice items to particular skill cells.
5. ***Collecting data from pilot subjects on items and item distributions.*** Correct answers and distractors (incorrect selections) may be modified from pilot data, with potentially multiple rounds in iterative development.

6. ***Field testing of CPS problems on designated samples that represent the population.*** This was completed in PISA in 2014.
7. ***Applying psychometric scaling of items, problems, and problem clusters.*** This is currently underway in PISA.
8. ***Collecting feedback from stakeholders on problems, items, and item analyses***. Reports are currently being prepared by ETS for OECD consideration.

Comparisons of Human-to-Agent (HA) and Human-to-Human (HH) in PISA

Two questions frequently arise from colleagues in PISA on the assessments with the agents and the limited number of interaction options associated with it. One question is how this approach compares with human-to-human CPS that takes place in computer-mediated communication. The second question is whether the multiple-choice questions are too limited compared with open-ended responses that usually occur between humans. These issues can, to some extent, be addressed by comparing HH with HA interactions and through systematically tracking the extent to which observations in the 12 skill cells yield scores that are complete, reliable, and valid within the given time constraints. The central issue is not whether the individual responses of humans are equivalent at each turn because they no doubt differ. Instead, the central question is how the completeness, reliability, and validity of the CPS competency scores may differ between the HH and HA approaches.

To address these concerns, OECD has commissioned a study to be carried out by a number of international experts in the field under the lead of the University of Luxembourg that compares HH and HA on some of the problem-solving scenarios employed in PISA 2015. In the study, one group will be solving the problems with HH, another group will use HA, and another group will use both HH and HA on several scenarios. This important study will assess how similar the CPS proficiency scales will be for the different conditions on both a rather general level of overall performance as well as on the level of the specific patterns of behavior in the assessment situation. This is a labor-intensive study with a complex design, because data in the HH condition will need to be coded in a comparable way to the original PISA 2015 tasks. Will the same individuals who are exposed to these two methods end up with the same or very similar CPS proficiency scales? If not, how do they differ? It is fundamental for the interpretation and reporting of collaborative problem solving in PISA 2015 to have empirical answers to questions such as these, and the study mentioned here will inform the reporting in PISA 2015 and will be the first that provides strong evidence of the similarities and differences of the H-H and H-A approaches with ties to international large-scale assessments.

ATC21S 2015 CPS Task Development

It was accepted as a basic principle in the development of HH assessment tasks for ATC21S that task development for collaborative problem solving ought not to vary from standard test development procedures. The major issue was the conception of what a test item would encompass in a CPS

assessment. The items were complex, and it was feasible that each item would have multiple solutions. More than one student would be involved in finding a solution, as the assessment was concerned with the process of collaboration as much as it was concerned with problem solution. The specifications focused on correlated strands as detailed in the conceptual framework (see Chapter 3). Each item was expected to yield multiple data points commensurate with the time allowed for the task. This was defined as the data efficiency of the task, and was a measure which could be compared to the data points yielded by an equivalent number of multiple-choice items, with one correct response yielding one data point per item. If, for example, the complex item occupied an average of 15 minutes for two students, a set of 10 data points might be expected to be yielded for each student, as would be the likely case for a 10 item multiple-choice question test.

In conceptualizing CPS, it was necessary to select the proper components and to define them in operational terms, as discussed earlier in Chapter 4. The domains (social and cognitive) identified by an expert panel were also defined, and hypothetical constructs underpinning them were proposed. Following conceptualizing of the construct and the hypothetical domains, a relatively standard approach to test development was followed.

The assessment tasks were expected to be suitable for students in the range of 11 to 15 years of age, which would cover upper elementary and middle school. The tasks needed to encompass both individual problem solving and collaborative problem solving, and their solutions needed to be digitally-based. Twelve principles were followed for the task development:

1. The tasks were expected to enable monitoring of individual student growth and development from lower-order skills to higher-order skills within the sub-strands of CPS.
2. The skills were expected to be teachable, measurable, and applicable to large-scale implementation.
3. The tasks should involve elements of ambiguity.
4. Tasks should engage students in ways that ensured they were dependent on one another for successful resolution.
5. No one student should be able to complete a task alone, and a task should not be solvable unless every student contributed to the problem solution. For this reason, each student should have unique control of a resource that was essential to the problem solution. That is, each of the students should access or control different but essential data and resources.
6. The tasks should enable teachers to interact and teach even while tasks are being completed by the students. The opportunity to teach and to promote learning while tasks are being undertaken is an important departure from many high-stakes, high-pressure forms of assessment. For this reason there was a required focus on learnable and teachable skills rather than personal attributes.

7. The assessment data should allow both formative and summative assessment decisions to be made to help improve both teaching and learning anywhere and anytime. Hence it was necessary to develop prototype tasks that were usable for a variety of purposes in education, with applications ranging from the individual student or collaborative groups to the classroom, the school, and the education system.
8. The range of tasks and data generated should allow student growth to be tracked across developmental progression levels.
9. There needed to be a process of automatic scoring of all tasks. It was not acceptable in ATC21S that the teacher be required to undertake the direct assessment observation and scoring of student involvement.
10. Background student activity data needed to be logged and collated and student activity monitored so that it could be coded, scored, calibrated, and interpreted in a way that was independent of the teacher.
11. The teacher's main involvement should be to administer the classroom arrangements for the assessment and to interpret reports with reference to the type of teaching instruction and intervention best suited to the student growth.
12. Feedback to students and teachers for purposes of formative intervention should involve a mix of task types, and a range of data should be provided to the teacher describing skills that students are ready to learn (not a score or a percentage or a grade).

TASK CHARACTERISTICS

Each task was constructed so that students could click, drag, and drop objects using the mouse cursor. There was a "chat box" for communication between partners, designed to ensure that students communicated through no other method (e.g. verbally across the classroom). Each task presented an instruction stem followed by a problem with tasks ranging in length and complexity from 1-8 pages. The tasks were designed to be recognizable as game-like puzzles at face value and contained aesthetically appealing graphics to attract and maintain student engagement. Students were assigned responsibility for different resources essential for the problem resolution. This obliged the students to work collaboratively, discuss the problem, and share resources. No student possessed sufficient information and resources required to solve the task individually. The difficulty of the tasks was manipulated by adjusting several of the parameters, such as number of problem states, constraints on object manipulation built into each task and described in the problem stem, the complexity of reasoning or planning required to guide the search and, finally, the configuration of objects and symmetries within the task (Funke, 1991). The tasks and their characteristics are described in detail in the research volume for the ATC21S project (Griffin & Care, 2015).

Coding

The approach to coding the log-stream data is provided in detail by Adams et al. (2015). For instance, it was taken that an example of social behavior was the presence of chat before an action took place. For example, a social behavior was assigned the code U2L001, which identified the task (U2), the action relevant only to the specific task (L), and the indicator number (001). When the student who performed the action was included, a letter (A or B) was added to the code (e.g. U2L001A). The purpose was to identify, categorize, and code every action and chat separately and in combination. Further developments in coding are offered by Scoular (2016), but the process has still to be refined as a set of template procedures that prompt task design. This is currently being investigated in a research study funded by the Australian Research Council (Griffin, Care, & Wilson, 2016).

Calibrating

Once the data were coded and recorded as item (indicator) level data, a calibration analysis was undertaken to identify indicators that did not behave in a manner required for the definition of the construct or the sub-strands of the construct. Fit to the Rasch model was used for this process. The calibrations were undertaken for one dimension (CPS), two dimensions (social and cognitive), and five dimensions (participation, perspective-taking, social regulation, task regulation and knowledge building). For the two and five dimension analyses, scored data sets from all 11 CPS tasks were combined. Misfitting indicators were removed, and it was then possible to obtain estimates of student ability and the relative difficulty of the indicators on each of the one, two, and five dimensions. Student ability was estimated using a subset of indicators from a selection of CPS tasks, so students were not required to complete all the tasks. Griffin, Care, & Harding (2015) present the details of the indicator calibration and the overall task calibration.

Development Process

Drafting

The first step in the development process involved a proof-of-concept scenario which sampled the skill-set described in hypothetical progressions formulated by expert panels (see below). Specific sub-skills of interest were identified and linked to contextual elements in the conceptual framework. Target student groups for each of the tasks were identified.

Paneling

The development process was guided by members of panels representing three specializations: collaborative problem solving, measurement, and classroom teaching.

The CPS expert panel consisted of specialists in problem solving, knowledge management, and computer supported collaborative learning. A measurement specialist acted as an advisory member. The task of this panel was to ensure that the proposed constructs were theoretically sound and that the evidence from student collaborative activity could be mapped onto developmental progressions.

They were also responsible for developing a framework within which the assessment tasks and teaching strategies could be conceptually based. The panel established hypothetical developmental progressions that represented measurement properties of order, direction, and magnitude of the construct (Wilson, 2009; Wright & Masters, 1983).

The measurement specialist panel had three roles. First, the panel scrutinized the CPS conceptual framework to determine its suitability for mapping the proof-of-concept tasks. Second, they reviewed the draft task materials and their potential to map onto the conceptual framework. Third, they examined the *data efficiency* of the tasks (the likely number of data points yielded given the time needed to typically complete the task).

A teacher panel was formed in each participating country. They reviewed materials and commented on the appropriateness of the tasks for the target student group and the likelihood that the tasks could discriminate between high- and low-ability students. They were also asked for their views about the learning and teaching potential of each task and requested to provide their qualitative feedback to national coordinators.

Cognitive Laboratories

The term *cognitive laboratory* describes the process of observing an individual (or, in this case, a student dyad) working through a task by "thinking aloud" and providing metacognitive data. The cognitive laboratory information was used by task developers to anticipate the kinds of coding that might be used to record the students' actions and reactions and eventually to code and score student performances.

Pilot Study

The small scale field pilot study focused on the administration, completion, and scoring of tasks. Field staff helped teachers observe the process and provide feedback. Ideally, a full class (up to 30 students) from each age level was used in each country. The documentation of the pilot study formed the basis of a field manual.

Trials

The major purpose of larger scale trials was to establish the psychometric properties of the tasks. It also enabled formative assessment reports to be designed and trialed with teachers and students. A delivery platform was required to enable collaborative interaction between students via the Internet. The platform was established with an architecture that allowed students to access the tasks in pairs, or dyads, for both problem solving and digital networks and real time scoring and calibration (Griffin & Care, 2015).

The activity log stream files recorded the student ID, the partner ID, the team or dyad identification, and relevant metadata for students undertaking the task. Every action was time stamped. The data identified the task and the web page of the task on which the students were working. The log file

data recorded whether the student had the role of student A or B, the event, the action taken by the student, any chat between the students, the start time, the finish time, progress within a task, and a description of the action or the actual chat. There was no attempt to interpret the chat itself, but interpretations were proposed for the process from the action and chat sequences.

Conclusion

The item development for any CPS assessment is very challenging for different reasons. Implementing the complexity of the framework, determining the item characteristics, and addressing the scoring issues are some of the necessary, but challenging steps. A very general decision is whether to use one or more computer agents as collaborator(s) or other student(s). The choice of the approach has strong implications for many of the other issues. For instance, what will the task environment look like? What will be the group composition? What collaboration medium should be used? How can the scoring be implemented?

The HH assessment approach is embedded into a less standardized assessment environment and offers a high level of face validity. A student collaborates with other students, so the behavior of both is difficult to control and the success of one student depends on the behavior of the other student and the stimuli and reactions that he/she offers. This has implications for scoring. How can the open conversation and large variety of stimuli be identified and utilized for scoring? In the HA approach, the assessment environment is more standardized. The behavior of computer agents must be preprogrammed, so the response alternatives of the item to which the student reacts needs to be limited to some extent, and every possible response of the student needs to be attached to a specific response by the computer agents' stimuli or event in the problem scenario. Such an approach ensures that the situation is rather standardized, and furthermore enables comprehensive scoring techniques since every possible turn in collaboration is predefined. Nevertheless, using the human-to-agent approach comes with the shortcoming of its artificial appearance and other conceptual questions, such as whether the CPS measured in this way can represent real-life collaborative problem solving. From a task design perspective, several distinguishing points arise from the different emphases of the two definitions. The two approaches are summarized in Table 5.1.

Table 5.1:

Comparison of two approaches to task development in collaborative problem solving

Issue	Human-to-Human (HH)	Human-to-Agent (HA)
Definition	Does not allow anything other than human involvement	Allows for two or more agents, with one being human and the others being computer agents.
Emphasis	Describing the process of collaborative problem solving	Competency of the human agent
Ability estimates	Interprets student ability measures as stages of increasing competence	Measures that enable a summative and norm referenced interpretation
Interpretation	Formative and criterion referenced	Point estimates for aggregation and analyses
Group composition	Group composition is central to the task design	Range of group compositions to take on different roles and capacities
Response constraint	Partners free to respond to partner or to initiate activities	Human interaction with several groups of virtual agents' response options are restricted
Group composition	Nature of the group composition has to be taken into account	Context is controlled as is the group composition
Competence estimation	Competence of each of the individual human participants can be identified	Estimation of competence of the interacting agent is not relevant
Development emphasis	Defines developmental progressions	Focuses on developing population parameter estimates from a sample of individual human subjects
Parameter estimation	Emphasizes the psychometric properties of the tasks	Estimating population parameters
Sampling	No need for probability sampling	Probability sampling mandatory
Platform	Platform that would enable broad adoption of the approach and materials	Platform not needed for system estimation
Scale	Not taken to scale	Taken to scale
Group effects	Allowance needed to be made for the student-within-team data structure and the interdependence among human participants	Not required for HA approach
Construct nature	Divides the complex CPS construct into two major components: social and cognitive.	Links the process of collaboration (consisting of three steps) with a derivation of Polya's four step process of problem solving

6 Scoring Collaborative Assessments

Introduction

This chapter discusses scoring collaborative problem solving measured in a NAEP setting. A unique aspect of a CPS task is that a test score could be generated by a group (e.g., students or students and computer agents) rather than by an individual student. In all current NAEP tests, each test taker independently contributes item responses, so this change could potentially require changes in many aspects of NAEP data collection and reporting. (This topic is discussed in more detail later in this chapter.) Next, we provide several examples of CPS tasks. We then discuss the types of data and interdependence in the data collected from CPS tasks. Only after that can we discuss scoring approaches and how specific analyses may inform them. The assessment of CPS skills is at an incipient stage and, therefore, many of the approaches presented here need to be further evaluated in subsequent studies. We conclude the chapter with a brief discussion about the validation of the claims associated with the CPS scores.

TASK TYPES THAT INFORM SCORING

PISA 2015 Collaborative Problem Solving

As mentioned earlier, the PISA CPS tasks are between a student and a computer agent, not between two students. Whether this constituted an actual collaboration is unclear, but a key point is that through the framework and assessment development PISA addressed central CPS scoring issues. The actual implementation of PISA 2015 (in contrast to the broader framework specification) relies mostly on multiple-choice questions at assessment points (rather than dialogue and other forms of student input), and for the Field Trial, simple sums of number correct were treated as an overall CPS score. At the time of this writing, it is unknown whether and how skill scores would be produced for the 2015 main study.

ATC21S Collaborative Problem Solving

To measure each of the elements, ATC21S process data captured in the log files are analyzed. The log files list each time-stamped activity undertaken by participants, along with contextual information. Analyses of log file data distinguish actions taken by the student from chats given by the student. Each of these actions and chats serves as the basic data which are then classified into the element categories. For example, taking actions or chats that can be interpreted as responding to another's contributions is scored as an indication of *interaction* (a kind of participation); actions or chats indicating ignoring the contribution of another is scored as a negative indication of adaptive responsiveness (a kind of perspective taking). Reaching a compromise with a collaborator is scored as *negotiation* (a kind of social regulation). Setting a goal for the task is scored as *goal setting* (a kind of task regulation element), and exploring elements of the task is scored as collects information (a kind of learning and knowledge building element).

The elements themselves, or perhaps more specifically, the element scores have been analyzed in such a way as to indicate six-level developmental progressions. For example, the lowest level, "1: Independent inefficient exploration," is characterized by no evidence of collaboration, and unsystematic and limited acknowledgements of the other. A middle level, "4: Cooperative planning," is characterized by indications of perseverance, successful subtask completion, awareness of the partner's abilities, and planning and goal-setting with the partner. The top level, "6: Strategic approach to problem via a collaborative process," is characterized by working collaboratively, systematically, and efficiently throughout the problem-solving process, tailoring communication, incorporating feedback, and resolving conflicts.

ADDITIONAL CPS TASK TYPES

ETS's Tetralogue

Like ATC21S, the purpose of ETS's Tetralogue was to jointly assess cognitive (science) skills and social (CPS) skills. ETS's Tetralogue combined a simulation-and-game-based science assessment that probes students using multiple-choices test questions, with a CPS task accomplished by a chat window. In the Tetralogue, student dyads worked collectively to make a prediction of volcano alert levels in a science simulation task.

The Tetralogue was designed to elicit student–to-student collaboration on one side and dyad-to-agents collaboration on the other side. The students were given the opportunity to first submit their individual answers, then to discuss and eventually revise them. The system then chose one answer from those submitted by the students and submitted it as the team response to the collaboration. Based on a review of computer-supported collaborative learning (CSCL) research findings (Barron, 2003; Dillenbourg & Traum, 2006; Griffin, Care, & McGaw, 2012; von Davier & Halpin, 2013; Halpin & von Davier, 2013; Halpin, von Davier, Hao, & Liu, 2014), the PISA 2015 Collaborative Problem-Solving

Framework (OECD, 2013), and evidence centered design principles (Mislevy, Steinberg, Almond, & Lukas, 2006), a Tetralogue conceptual model was developed that documents a matrix of individual and social skills involved in collaborative problem solving, which provides a basis for designing assessments of CPS (see Liu, Hao, von Davier, Kyllonen, & Zapata-Rivera, 2015).

The data from the interactions of the two student collaborators are rich time series and can be modeled statistically using appropriate methods. Some of these approaches have been studied for educational assessments by von Davier & Halpin (2013) and Halpin, von Davier, Hao, & Liu (2017) and for team performance in other contexts, such as military, sports, and games (Soller & Stevens, 2008; Halpin & von Davier, 2017; and Littman, 2000, respectively). The models used in these papers for the analyses of the process data from CPS are from the family of stochastic processes. One of the most important features of the CPS process data that are modeled in this stochastic framework are the interdependencies among the team members' actions over time.

A set of other testing instruments were administered together with the Tetralogue CPS task: a science multiple-choice test, a personality test, a background questionnaire, and a post-test survey to evaluate the collaborative experience.

CMU-MIT Collective Intelligence Battery

The CMU-MIT Collective Intelligence Battery (Woolley, Chabris, Pentland, Hashmi, & Malone, 2010) consisted of tasks drawn from the McGrath Task Circumplex, a taxonomy of group tasks varying in the nature of the collaboration between team members required. Tasks included a brainstorming task, a matrix reasoning task, a moral reasoning task, a shopping trip planning task, and a group typing task. The purpose of creating the battery was to conduct a study on the existence of a "collective intelligence"—analogous to general intelligence for individuals. In two studies with several hundred participants, groups of two to five members performed the various collaborative tasks. Evidence for a collective intelligence factor was shown—groups differed systematically in how well they performed the CPS tasks (i.e., some groups were consistently better across tasks compared to other groups). Interestingly, individual cognitive abilities were not strong predictors of collective outcomes. However, one ability was correlated with collective output—a social sensitivity measure. A process measure—conversational turn taking—also predicted group performance. Other measures believed to relate to group success—group cohesion, motivation, satisfaction—were not found in this study to predict success.

The key lesson from the Woolley et al. (2010) study, and why it informs a CPS task for NAEP is the evidence it presents that performance on a CPS task can be very different from performance on a typical cognitive task, such as the ones in NAEP. The implication is that collective output, as well as any individual performance measure, may be an important variable to measure in its own right.

Table 6.1:

Summary of task characteristics

	Number and Type of Tasks	Communication mode	Collaboration	Process data	Item Types
PISA 2015	4 computer-based tasks	Chat	Human-to-Agent	Minimal/logfile. The chats are deterministic and have been automatically scored. The timing information is the main variable for the process.	Multiple-choice (MC) or partial credit (PC) items
ATC21S	11 computer-based short simulation tasks	Chat	Human-to-Human	Detailed/logfile. The chats have been minimally scored by human raters	MC or PC items
Tetralogue	Computer-based; one long science simulation	Facilitated chat	Human-to-Human	Detailed/logfile. The chats and actions have been scored by human and AI raters	One long science simulation with embedded MC items
Collective Intelligence Battery	5 complex tasks	Face-to-Face	Human-to-Human	Video/audio. Scored by human raters	Complex tasks

Data from CPS Tasks

Students in collaborative settings talk, negotiate, hypothesize, revise, and respond, both orally with gestures and online with chats and emoticons, acronyms, and so on. All of these seem to matter within the context of collaboration. Data from CPS tasks can be characterized as individual and team (collective) outcome data and process data. Examples of outcome data are the correct/incorrect assessment of an action or task at the individual or team level. Process data offer an insight into the interactional dynamics of the team members, which is important both for defining collaborative tasks and for evaluating the results of the collaboration (Morgan, Keshtkar, Duan, & Graesser, 2012; Morgan, Keshtkar, Graesser, & Shaffer, 2013).

The process data from CPS consist of time-stamped sequences of events registered in a log file that describe the actions and interactions of the students and of the system. This kind of process data has so far been the grist of educational data mining (EDM; Baker & Yacef, 2009; Romero & Ventura, 2005; Kerr & Chung, 2012; Kerr, in press). Next, the outcome and process data and their role in the CPS assessment are discussed.

Process Data in Human-to-Human Collaborations

In the context of learning and assessment technology, time-stamped activity events can represent the minutiae of students' activities, right down to the level of where the pointer is located on the screen. A challenge in analyzing log file data is determining the meaning of individual actions and chats. There may be some process variables that are relatively easy to measure, such as the participation level of each team member and turn taking. However, beyond these kinds of variables, interpreting actions and chats may be much more complex because of the dynamics and the sheer volume and complexity of data generated in log files.

Regarding dynamics, consider that in collaborative problem solving, interactions will change over time and will involve time-lagged interrelationships. If there are two people on a team, the actions of one of them will depend both on the actions of the other team member and on his or her own past actions. The statistical models used need to accurately describe the dynamics of these interactions. These dynamics, which are defined by the interdependence between the individuals on the team, could offer information that could be used to build a hypothesis about the strategy of the team. For example, by analyzing the covariance of the observed variables (the events), one might hypothesize that an unknown variable, such as the team's type of strategy, explains why the team chose a particular response and avoided the alternative.

Regarding the volume and complexity of interpreting actions and chats, note that collaborative interactions in computerized educational environments produce data of extraordinarily high dimensionality (often containing more variables than people for whom those variables are measured). Extracting key patterns from the noise in such data is crucial, not only to reduce the dimensionality but also to identify relevant features of student performance in these complex tasks (Kim et al., 2008).

The research programs described earlier resulted in observable process data that could be directly linked to the students' performance/outcome. For example, see Hao, Liu, von Davier, Kyllonen, & Kitchen (2016) where the relationship between the CPS process attributes, such as negotiation, is an indicator of a successful outcome.

Process Data in Human-to-Agent Collaborations

The HA collaborations in PISA circumvented the challenge of analyzing the HH dialogs and dependent actions by having multiple-choice options for each human turn in the chat. There were also a limited number of action alternatives at points when the human was expected to perform a nonverbal action (e.g., click values in the control panel, click the button "Tryout conditions"). In essence, the multi-turn and multi-action interactions were converted to multiple-choice scoring events. The multiple-choice approach also can be engineered to guarantee a coherent computer response to each alternative available to the human by developing a map for conversational threads

and implementing convergence zones (states that all humans reach during the course of completing a problem) to reduce the solution space. This approach is described in Chapter 4.

A more graded scoring system could have distractor items receiving partial credit for a cell or have a correct answer contribute to multiple cells. If open-ended items in human chat are desired, then a contribution of a student in a conversational turn needs to contribute scores to one or more of the cells.

SCORING A CPS TASK

Scoring Traditional Multiple-Choice (MC) Items

PISA 2015 leaned heavily on a limited number of options for the student to choose from at each assessment episode in a chat or actions during the exchanges with agents and the computer system (the multiple-choice approach). The response options in each assessment episode targeted one or more of the 12 skills measured in PISA. Therefore, standard psychometrics such as IRT methodology could be applied to these data. There are approximately a dozen of these MC items in a particular problem to be solved in PISA, but the number of choice points could potentially increase in number. Differential item weighting, partial credit scoring, and multidimensional scoring are all possibilities although none were implemented in PISA.

The PISA 2015 assessment has each student solve approximately five problems in two 30-minute sessions. Each problem has a fixed set of MC items that span a specific convergence zone in the interaction. The MC items need to cover all 12 cells in the construct matrix of the PISA Framework (see Table 3.1). Performance in the 12 cells determines the competency measures for CPS. These cells in the matrix can, of course, be weighted differentially.

It is important to construct MC items so that sophisticated guessing cannot give a student any advantages. For example, it is important for the student to take the initiative and control of a task in some situations, but not others. Sometimes a student should disagree with or not follow a team member, but not always. The correct answers to an ensemble of items needs to be nuanced with context sensitivity and to be impervious to gaming the system or indiscriminant application of simple strategies.

Scoring Chats: Dialogic Scoring

Open-ended student responses require advances in computational linguistics in order to score them for the 12 cells specified in the PISA 2015 framework (OECD, 2013; Table 6.1, Chapter 6). As a note, PISA could not accommodate automated responses because of the constraint of accommodating multiple languages. However, NAEP would be in English, a language in which there have been impressive advances in computational linguistics. The following steps need to be accomplished in an automated analysis of open-ended language in student turns.

1. Segment the language within a student turn into speech act units.
2. Assign each speech act unit to a category, such as question, request, acceptance, denial, short positive response, expressive evaluation, and so on.
3. Assess the extent to which each student speech act (or turn) contributes score values to each of the cells in the framework.
4. Create CPS competency measures from the score values in the cells. The competency measures can be attributed to individual students, the processes of team members interacting, or the achievement of the group goals.

In PISA 2015 the student clicks on multiple-choice options on chat turns or on alternative actions that they can perform. These decisions can be mapped onto the 12 cells of the PISA matrix and steps 1 and 2 can be skipped.

One viable approach is to have human-to-agent interactions with open-ended human responses in natural language, as opposed to selection of MC options. The chat maps could be set up in the same way as the PISA CPS 2015, with convergence zones. However, the human input would be natural language rather than the selection of an option from a menu. Advances in computational linguistics in the English language and culture are sufficient to seriously address this possibility. Also, agents individually or in pairs (or system states) can barge in and end one convergence zone and start another.

Listed below are some breakthroughs in computational linguistics that make it feasible to consider open-ended responses:

- Student conversational turns are typically short, the vast majority being one or two speech acts. This makes it easier to segment the sequence of words into speech acts units.
- Speech acts can be classified into categories (e.g., question, request, statement/assertion, agreement, etc.) with a moderately impressive accuracy.
- The content of students' assertions can be matched to expectations (i.e., sentoids in a rubric) to see whether their language matches what the computer expects in a turn. Techniques from computational linguistics (CL) can do this with a high degree of accuracy in many but not all knowledge domains. Latent semantic analyses, Bayesian analyses, and regular expressions provide foundations for evaluating student-language-with-expectation semantic similarity matches that are equal to human evaluations of similarity.
- There are different categories of discourse expressions that are frozen, vernacular, or very specific, with a limited number of ways to express them. Examples are "I don't know," "okay," and "What's next?" These expressions can be detected with a high degree of accuracy by CL technologies.

- Discourse patterns have been identified for a number of common communication functions, such as the grounding of referring expressions, giving feedback to peers, answering questions, expressing disagreements, giving hints, providing help, and so on. These patterns can be detected with moderate accuracy through advances in the computational discourse sciences.

ETS's implementation of the Tetralogue opted for unconstrained open-ended responses of the two human collaborators in natural language. The analysis of the language features of the conversation was conducted. An automated classification scheme was developed to classify their conversations into some predefined social skills (Liu et al., 2015) that were based on the PISA 2013 CPS framework (Graesser & Foltz, 2013), but expanded to include other factors that were found to be relevant for CPS. In the classification, there were 39 subcategories corresponding to four main categories of CPS skills: sharing resources/ideas, assimilating and accommodating knowledge, regulating problem-solving activities, and maintaining positive communication. The goal was to examine language features of participants' communication and explore whether these features could allow for automatically detecting participants' CPS skills with some basic natural language processing techniques. To develop an automated classification engine, we prepared a training and a validation dataset with known "ground truth." Two human raters scored the responses from dyads into the categories of social skills mentioned above.

TagHelper (Rose et al., 2017) was used for classifying sentences. A *feature* can be thought of as a unit of dialog that is computable. There are a number of language features available for extraction in the TagHelper, such as the unigram, bigram, part-of-speech tagging of bigram, sentence length, punctuation, and word stem. Meanwhile, it also provides a selection of machine learning algorithms to map the features to the target labels. For the Tetralogue data set, the language features with only unigram and bigram, together with the Support Vector Machine (SVM) algorithm provide the best classification result based on a 10-fold cross validation. Currently the manual coding scheme is being refined to achieve higher inter-rater reliability (.7). These approaches are described in Rosé, Wang, Cui, Arguello, Stegmann, Weinberger, & Fischer (2008) and in Wang, Hao, Liu, Chen, and von Davier (2015).

The reliability of an assessment with CPS tasks will depend on the number of "items" (opportunities to display a collaborative behavior), the variability of partner's attributes and the variability of the task type. These are all topics for further research.

Analyses That Can Inform Scoring Approaches

Scoring collaborative problem-solving tasks is still very much a research project. More needs to be done before we can confidently make inferences about collaborative skills based on item responses. In this section, we describe research that may inform scoring decisions. Outcome data (correct/incorrect/partially correct data) can be analyzed with classical test theory (reliability, correlations, biserial correlations), factor analysis, and IRT. Process data may also be analyzed in this way under specific circumstances and may also be analyzed using data mining tools and time series methods.

Constrained process data (where chats and actions are interrupted by an agent or the system to get back onto a previously anticipated path) may be simplified or aggregated and then classified by a human scorer into predetermined categories—this is the kind of approach that could be taken in NAEP, for example. The human scorer could treat actions and sequences of actions in the same way a human essay scorer treats essays: He or she uses reasoning skills to make inferences about the intents and skills of the test taker. There are issues about rater effects, but these can be handled with conventional approaches, for example to determine interrater agreement, provide rater training, and provide a subset of tasks and data for calibration purposes.

There are various analysis approaches we could take to investigate the measurement quality of the CPS tests.

- Analyzing the distribution of assessment scores for the attributes in Table 6.1 (or other identified skills that will be considered for NAEP). This includes ensuring that there is an adequate number of observations for each attribute, and assessing interrater agreement in interpreting responses to particular multiple-choice items as evidence for a particular attribute.
- Conducting distractor analyses for multiple-choice responses, including computing item response-total correlations or computing item characteristic curves (parametric or nonparametric). Keys and distractors may be modified from such pilot data.
- Field testing of CPS problems on designated samples that represent the population.
- Applying data mining tools to identify different patterns and strategies in problem solving.
- Applying stochastic models to dependent data over the course of collaboration.
- Applying psychometric scaling of items, problems, and problem clusters.
- Computing inter-item correlations, and factor analyses of the resultant matrix.

If the data result in a long time series for each team member, data mining or time series models may be appropriate. Data mining does not have a long history in education or psychology because, until recently, educational and psychological data were not often of high enough dimensionality to require such techniques. However, these techniques have been used for decades in fields where data with high dimensionality have long been the norm, such as finance, marketing, medicine, astronomy, physics, chemistry, and computer science, and most recently in computer-supported learning (Romero, Ventura, Pechenizkiy, & Baker, 2010). The purpose of data mining techniques is to reduce the dimensionality of the data by extracting implicit, interesting, and interpretable patterns. In this scenario, different clusters of patterns of responses may be assigned different scores.

Other methods that can be considered on these rich data that exhibit time dependence at the individual level are multivariate stochastic processes, time processes, and dynamic models. Some of these models can accommodate analyzing the interactions among multiple team members; other models are more appropriate for identifying collaborative strategies of collaborative events (Soller &

Stevens, 2008; von Davier & Halpin, 2013). Most of the modeling approaches for these process data do not come from educational assessment, however, so they must be adapted from other fields.

From the perspective of psychometric theory, Bayesian Belief Networks (BBNs) are by far the most useful means that have been used so far for modeling student knowledge that is demonstrated during a complex task (Russell & Norvig, 2003; Levy, 2014). BBNs model the probability that a student has mastered a specific knowledge component conditional on the sequence of responses given to previous elements of a task. BBNs have long been applied in simulations and games to represent student knowledge and, thereby, guide the activities of the tutoring system and in the design of complex assessments (Mislevy et al., 2002; Shute, Hansen, & Almond, 2007; VanLehn & Martin, 1998; Vomlel, 2004); therefore, BBNs are an obvious methodological bridge between CPS and traditional psychometric theory. However, the practical implementation of BBNs often requires highly simplifying assumptions, and, as with traditional models, they have not been adapted to represent the knowledge of multiple individuals simultaneously. Therefore, the BBNs may be applied to score the team's process and outcome data first, while extensions of the BBNs are being developed.

A Statistical Representation of Collaboration

Von Davier and Halpin (2013) modeled collaboration as statistical dependence among the activities of two or more individuals. This approach also fits nicely with traditional theories of teamwork that distinguish between independent and interdependent teams (Thompson, 1967). They proposed to measure the degree of interdependence demonstrated by the activities of the individuals in a team by using the Kullback-Leibler divergence (KL) of the marginal distributions from the joint distribution (von Davier & Halpin, 2013). The interpretation of KL in the context of collaboration is intuitive. If the KL is zero, then we observe an independent team. When the KL is positive, some interdependence is exhibited among the activities of the individuals, with larger values indicating more interdependence (i.e., a greater divergence from the model of independence).

Von Davier and Halpin (2013) defined an outcome as a function of the complete time series. Most simply, if the activities recorded are correct responses to the components of a CPS task, we could define the group's total score on the task as the sum of the outcomes from all team members to all task components and compute its expected value accounting for the degree of interdependence. For example, for an independent team, the expected outcome is simply the sum of its parts.

A productive collaboration can then be defined as one for which the expected outcome is greater than the sum of its parts, and an unproductive collaboration would have an expected performance worse than that of the independence model. A similar approach can be applied to other collaborative outcomes. Instead of sum scores, it will be generally advantageous to have a psychometric model for the entire response pattern, for instance, an IRT model or a BBN.

In modeling the *processes* of collaboration, we are concerned about describing the statistical dependence exhibited by the activities of groups of individuals. In modeling the outcomes of collaboration, we are concerned with judging the performance of a group relative to what would be expected from the individual group members had they not collaborated. Von Davier and Halpin's (2013) model incorporates both of these aspects. In CPS, we are concerned with how the probability of an individual's activities changes over continuous time as a function of previous activities. A point process is a model for isolated events occurring in continuous time, with familiar examples including the Poisson and renewal processes. The passage to point processes can be made by considering the temporal structure of the activities during the collaboration. With the aid of the Hawkes model Halpin & De Boeck (2013), it may be possible to identify different types of events during the interactions based on the interdependence in the data, such as autoregressive events, spontaneous events, or cross-dependent events. This classification may result in a "collaboration index" (Halpin, von Davier, Hao, & Liu, 2014) which may be considered in the scoring process. Von Davier & Halpin (2013) also combined the Hawkes model applied to process data with the outcome or scored subtasks during the CPS interaction, which resulted in a marked Hawkes model. More research is needed for the application of these models to score CPS tasks. Nevertheless, a statistical latent-variable-based framework for the CPS skills will allow for the psychometric requirements of the scores to be met: reliability and standard errors will be directly available for scores derived using marked Hawkes model.

Conclusion

This chapter discussed several scoring methods that may be appropriate for CPS tasks and discussed in more detail the research around the inclusion of the process data in the scoring process and the differences between the human-to-human and human-to-agent collaboration. The score reporting considerations for these approaches are further discussed in Chapter 7.

As the research on tests with CPS tasks progresses, so does the need for validity considerations. Coordinated research studies are needed to support the claims of the scores for various tests. The validity research should be tied to claims about collaborative problem solving and evidence in support of those claims (Kane, 2013). (This is discussed further in Chapter 7 for the score reporting specific to NAEP.) The validity studies may take a few years to be properly conducted as all will require additional data collected from students from the same target population.

As illustrated in this chapter, there is more than one way to score an assessment with CPS tasks. Test scores depend on the test use and on the choices made about test design and analyses. These components affect one another during the assessment development, so the evaluation of the scoring mechanism in isolation from the rest of the system is only a piece of the larger validation argument.

7 Reporting

Reporting is the most publicly visible component of a NAEP assessment. What is reported from an assessment plays a major role in how the assessment is received both by the general public and by the more direct stakeholders, including teachers and school, district, and state personnel. NCES produces the initial release NAEP reports in collaboration with the National Assessment Governing Board, and both Congress and White House staff have referenced results from NAEP reports in their education briefings; thus reporting is important for policy implications arising from the assessment.

Current NAEP Reporting

A major issue NCES must address for all its reporting is obviously what to report on. Existing NAEP reports may be useful as guides for envisioning how results from a collaborative assessment might be reported.

NAEP now presents its reports online, and in addition, provides dashboards;[1] data tools for teachers, researchers, policymakers, and secondary analysts; news releases; and sample questions. The online reports themselves, such as the biennial mathematics and reading reports, feature trend comparisons by scale scores and percentages at or above the NAEP Achievement Levels (*Basic*, *Proficient*, and *Advanced*). Results are typically disaggregated by various student variables including gender, race/ethnicity, school type and location, students with disabilities, and English language learners. Trends in score gaps by gender and race/ethnicity are also featured prominently in NAEP reports. Data are further disaggregated in reports by state and by trial urban districts. There also are survey questionnaires for each assessment that are administered to assessed students, their teachers, and school leaders. These questionnaires typically address issues such as courses taken, perceived difficulty of courses, attitude towards the subject, use of tools (e.g., calculators, computers) for homework and tests, after-school tutoring, perceived effort on and importance of the NAEP assessment, as well as general questions that reflect demographics (e.g., race/ethnicity, parental education, home possessions, family language) and school related activities (e.g., amount of homework, absenteeism).

[1] NAEP dashboards highlight the percentage of students at or above a proficient level for 4th, 8th, and 12th grades, across civics, economics, geography, reading, science, history, and writing.

Special, new, or novel assessments—such as the new technology and engineering literacy assessment (TEL)—focus particularly on helping the user understand the nature of the construct assessed and tend to highlight the more innovative nature of the assessment by presenting and discussing specifics of the assessment content. In the case of the TEL report released in May 2016, several of the scenario-based tasks (SBTs) that made up a significant portion of the assessment content were featured prominently and in a variety of ways on the TEL website. The TEL report may be a particularly relevant model for how collaborative problem solving results could be reported.

What to Report: Elements Reflecting Construct Definition

COLLABORATIVE PROBLEM SOLVING

One of the major substantive issues for a collaborative problem-solving task would be whether to report collaborative skill, problem-solving (content) skill, or both. A related issue is whether to report on collaborative or problem-solving skill at the individual student or team level. Although NAEP is a group score program (Mazzeo, Lazer, & Zieky, 2006), the reported score statistics are based on individual responses, and the student is the unit of analysis. In collaborative problem solving there are both individual and team responses. An individual response is the action the individual takes on his or her own; a team response is an action the team takes as a group, regardless of who actually issues the response.

It is possible to separate them (Griffin, 2017; Hao, von Davier, & Kyllonen, in press). For example, Hao et al. (in press) discuss a collaborative science assessment, where a science achievement score is computed from the final responses on problems, but collaboration scores are computed based on chats (conducted in a chat window) and related interactive activities. Student-level data can be obtained from the student's initial response given before collaboration begins. Griffin (2017) discusses getting separate content and social scores.

CONTENT

As noted above, NAEP conducts assessments in numerous domains, such as mathematics, reading, science, social science, writing, and technology and engineering literacy (TEL). A collaborative problem-solving assessment can be embedded within those domains similar to the way Hao et al. (in press) did so in science. The basic strategy is simply to allow teams (of two or more students) to work together on a problem from a domain through an interface that enables communication via a chat window, and captures all keystrokes and mouse movements. Students still can be asked to enter responses on their own, as a team, or both (in Hao et al. [in press], both individual and team responses are elicited). For this strategy, an assessment could be developed from existing NAEP content frameworks and specifications for a particular domain, and a collaborative component could

be added to the assessment. This strategy would work best with scenario-based tasks, such as the TEL assessment (NCES, 2016), and the 2009 Interactive Computer Tasks (ICTs) Science assessment (NCES, 2012).[2]

Alternatively a new assessment domain, not already covered in NAEP content frameworks, could serve as the basis for the collaborative problem-solving assessment. An example here is the PISA collaborative problem-solving assessment (He, von Davier, Greiff, Steinhauer, & Borysewicz, in press), which is based on a specific collaborative problem-solving framework (OECD, 2013). In the OECD framework, four problem-solving stages (exploring, representing, executing, reflecting) are crossed with three collaborative skills (understanding, action, organization) to create a 12-cell CPS skills matrix (see chapter 3). Griffin's (2017) CPS framework is similar. Griffin crosses three social components (participation, perspective taking, social regulation) with two cognitive components (task regulation, knowledge building). Either of these could serve as the basis for test specifications for assessment development, and could provide potential reporting categories (e.g., scores reflecting skills at the cell level, such as participation in knowledge building). Empirical analyses and findings could suggest the degree to which individual scores at the cell level could be justified for reporting. It is important to note that in general there are many dimensions in social and emotional skills, perhaps more so than in general cognitive skills (Kyllonen, 2017).

Regardless of whether a cognitive-by-social skills breakdown could be accommodated empirically, it could still be the case that, in reporting on collaborative skills, content skills would simply serve as part of content specifications to ensure full construct coverage across domains. However, content skills would not be part of the reporting per se. Instead, under this scheme content skills would be construct irrelevant variance.

CONDITIONS/CONTEXTS

The framework documents associated with collaborative problem solving will specify the degree to which the conditions and contexts in which CPS is measured are part of the construct (e.g., "collaborative problem solving involves communicating in technology-mediated environments") or that the conditions are intended as valid variations of performance that support the intended interpretation (e.g., "interaction with avatars is equivalent to interaction with humans in CPS"). Such contexts could also be part of a design framework with CPS being tested across various contexts (e.g., human-to-human and human-to-agent). Such assumptions would be articulated both as part of the framework documents and as part of reporting. Besides technology-mediated environments and human-to-human vs. human-to-agent contexts, other conditions would include mode of delivery, degree of support, multiple-choice vs. short-answer response formats, and other test administration conditions.

[2] NCES is gradually introducing scenario-based tasks in other subjects. They are introducing 2 per grade in reading and math, beginning in 2019. Science may have about 5 to 7 per grade by 2017.

What to Report: Assessment Grades and Years

NAEP is generally administered to students at grades 4, 8, and 12, although some assessments are limited to specific grade levels. For example, economics has only been administered to twelfth-graders, under the assumption that economics courses are not taken by students in grade 8 and below. The initial technology and engineering literacy (TEL) assessment (winter 2014 administration) was administered only to eighth-graders, although future plans may include additional grade levels.

A model similar to the TEL rollout could be followed for collaborative problem solving, introducing it first for grade 8, followed later by administrations at grades 4 and 12. Most CPS assessments (e.g., Griffin, 2017; OECD, 2013) have been administered to students in the middle school to early high school years, therefore CPS large-scale assessments are best understood at this grade level. There would be additional challenges for grade 4, particularly in communicating through chat windows.

Collaborative problem solving is an assessment similar to the TEL assessment in that it is not explicitly tied to the curriculum, and it may not be as central to policy interests for some as mathematics and reading assessments might be. Therefore, it would likely be administered less frequently than mathematics and reading assessments, and more in line with the rate at which a TEL assessment might be administered.

Score Structure for Reporting

SCALES AND SCALE SCORES

A collaborative problem-solving framework such as the Griffin (2017) or OECD (2013) frameworks specify a number of categories (content/problem solving by collaboration), each of which potentially could represent a separate reportable skill. Assessment development based on such frameworks, combined with empirical analysis of findings (e.g., factor analysis) could provide evidence for whether such skills are separable (discriminant validity evidence) or not. It may be that a general collaborative problem-solving skill could be reported (essentially averaging over the different cell categories), or sub-category CPS skills, such as collaborative participation, perspective taking, and social regulation could be reported on separately.

NAEP assessments use a variety of strategies for computing and reporting scores. For NAEP's core subject-area assessments, such as mathematics (NCES, 2015), responses are analyzed using item-response theory and are reported on a 0 to 500 (or 0 to 300 for 12th grade) scale. The scale is a composite scale, and is made up of separately scaled components, such as number properties, geometry, and algebra. The scale is computed in a base year, combining 4th- and 8th-grade scores, and subsequently, new data collections are linked to the original cross-grade base-year score. Score disaggregation by the major reporting categories, and background questionnaire summaries are also reported. For science, the main science scale is a univariate scale—not a composite of the weighted subscales.

For science hands-on tasks (HOTs) and interactive computer tasks (ICTs) (NCES, 2012), scores were reported as follows: (a) item percentage correct (of items attempted, so as not to penalize for items not reached) were reported by grade and by gender, race/ethnicity, and income (eligibility for the National School Lunch Program); (b) path analyses of students' patterns in working through the task was reported; and (c) example problems were presented with the percentage of students answering them correctly. This form of scoring and reporting for ICTs and HOTs was done because of the limited number of tasks that were developed and administered.

In thinking about how collaborative problem solving might be reported, an approach of initially limiting analyses then over time providing more elaborate approaches may make sense. For example, initial reporting might be limited to percentage correct disaggregated by reporting variables, combined with presenting example problems, and perhaps some process variables. With experience with the assessment, it may be appropriate to conduct item-response theory analyses and report on a cross-grade base-year scale based on a composite score. A background questionnaire along the lines of what was developed for the TEL assessment could also be administered.

ACHIEVEMENT LEVELS AND SCORE PRECISION

For most NAEP assessments, achievement levels (*Basic*, *Proficient*, and *Advanced*) are set through a standard setting process overseen by the National Assessment Governing Board. The process involves a committee of experts, educators, and other representatives who decide on minimum standards for the respective achievement levels. For example, for a recent twelfth-grade NAEP reading and mathematics assessment a modified bookmark standard setting procedure (Cizek & Bunch, 2007) was conducted to reflect academic preparedness for the various areas of study (NAGB, 2012). Benchmarking studies (e.g., against college freshman) are also conducted occasionally.

NAEP reports scores by percentage achieving the various achievement levels disaggregated by the major reporting variables (e.g., gender, race, income). NAEP regularly reports percentile scores, and examines trends in those percentiles across years, enabling inspection of score trends by subgroups. Additionally, NAEP reports on the kinds of questions that students at certain levels are more likely to answer than students at lower levels. These procedures of standard settings, benchmarking, trends, and performance levels could also be followed for a CPS assessment. There would be complications due to the nature of collaborative work, but such complications could be addressed through research prior to a study.

Because of the complex design of NAEP assessment administrations, computing reliability or score precision is not straightforward. One reason is that it is a group-score assessment, meaning that scores are reported back at the group rather than individual test taker level (Mazzeo, Lazer, & Zieky, 2006). A collaborative assessment could be even more complex because the unit of analysis would not necessarily be the student, but the group of students. This could necessitate a design different from the current NAEP design.

Reporting Groups

The major reporting groups for NAEP are gender, race/ethnicity, type of location, parents' education level, eligibility for the National School Lunch Program, type of school, and region of the country. Such reporting groups could be retained for collaborative problem solving, although a complication is that these reporting groups pertain to individual test takers, whereas collaborators could belong to more than one reporting group. Research would have to be conducted to determine how best to address this situation. For the NAEP mathematics, reading, science, and writing assessments, reporting is also done at the state and trial urban district level, in addition to the national level, and such a breakdown could be envisioned for a CPS assessment, if program resources support such an administration design and sample.

Score Reporting – Interpretive Supports

With large-scale assessments such as NAEP or PISA, it is common to include performance level descriptions in the form of items that can be solved by students at a certain ability level (with some probability, e.g., 50%). Such a process could be implemented for collaborative problem solving, although again, due to the collaborative nature, it may be more challenging than is typically the case with mathematics or reading items.

Another aspect to NAEP (and PISA) reporting concerns relationships between assessment scores and other variables. One way this is done is through the reporting categories (see previous section), such as gender, race/ethnicity, and so forth. Another way this is done is through cross-tabulation with questions from the survey questionnaires. For example, PISA presents scale scores for students who provide different answers to questions pertaining to self-concept, math anxiety, self-efficacy, and the like. Something similar could be done for collaborative problem solving. Again, the complication is that in reporting on team problem solving, the members of the team could be (and are likely to be) in different categories with respect to such contextual or demographic variables. Woolley et al., (2010) measured individual differences variables to develop a model that predicted success in collaborative problem solving, finding that individual problem-solving skill was not as strong a predictor as were emotional intelligence and being female in predicting collaborative success. Research similarly could be conducted on collaborative problem solving. An important question is what variables would be important to include in a questionnaire. An expert group could be assembled to weigh in on such an issue.

Score Report Design

NAEP provides various reports—individual state reports, national reports, multi-state reports, trial urban district reports—for various audiences. A similar model could be followed for collaborative problem solving. Both PISA and NAEP provide good examples of report design in both paper- and online-based reports in terms of narrative and graphical elements, key messages, and so forth (NCES, 2012; OECD, 2013). NAEP and PISA also present opportunities for secondary data analysis

through the NAEP data explorer as well as by opportunities for scientists with data analysis training. There are other examples available in the states (e.g., the Florida Department of Education for an assessment-centric example: http://www.fldoe.org/accountability/accountability-reporting/interactive-reporting/; the Kentucky Department of Education for an example that includes assessment and many other data elements: http://applications.education.ky.gov/SRC/).

Conclusion

NAEP and PISA provide excellent models for score reporting on national and international assessments. Regular NAEP assessments, such as mathematics and reading, provide good models for design, analysis, scoring, and reporting using probability (cluster) sampling, balanced incomplete block designs for administration, and item-response theory for scoring, scaling, and linking that enable comparisons across time. Unique NAEP assessments, such as the science hands-on task (HOTs) and interactive computer tasks (ICTs), and the new TEL tasks provide other models for reporting. PISA 2015, the score report that was released in December 2016, provides a reporting model for collaborative problem solving. However, it is important to note that PISA collaborative problem solving will be very similar to individual problem solving as only one person collaborating with an agent is involved. If NCES were to create a collaborative task involving two persons, then important research would have to be conducted and decisions made concerning how to report on the collaborative output given that two (or more) persons would be contributing to that collaborative output. The two persons might differ with respect to background characteristics and other reporting variables, which would make reporting more complex than the usual single-person reporting.

8 Implications for NAEP

As the only federally-funded national assessment of students in elementary-, middle-, and high-school, NAEP has long played a unique and highly prominent role in American education, and has been a leading monitor of student academic progress and a key source of innovation in educational assessment.

NAEP's role as a monitor is evident from the press attention its reports receive. As one analysis of media coverage of test results notes:

> NAEP is the only measure of its kind to produce nationally representative and continuing results of student achievement in the country. It is widely respected by educators and researchers alike for its technical rigor and substantive contributions to the understanding of student achievement....In short, this test has an unusual degree of credibility and importance, which should translate into critical attention to its results and a high level of visibility in the media (Ogle & Dabbs, 1998, p. 87).

This level of credibility and prominence reflects several features of NAEP's design. It carefully draws samples of students to provide an accurate picture of student achievement, both nationally and in states and large urban districts. It does not provide results at the student, classroom, or school level, and thus is much less susceptible to teaching to the test or other factors than tests with higher stakes attached can experience. It also remains stable over time and thus can provide trend data that many other tests cannot provide.

Another factor contributing to NAEP's status as the premiere monitor of student academic performance is the way its assessment frameworks are developed. The National Assessment Governing Board, which develops NAEP frameworks, uses an extensive process that involves the following:

- widespread participation and reviews by educators and state education officials;
- reviews by steering committees whose members represent policymakers, practitioners, and members of the general public;
- involvement of subject supervisors from education agencies;
- public hearings; and
- reviews by scholars in the field, by the NCES staff, and by a policy advisory panel.

In that way, NAEP's assessments reflect a general consensus in the field about what should be tested.

In addition to serving as a monitor of student achievement, NAEP is also widely recognized as a leader in assessment innovation. Some of the innovations have been methodological; for example, NAEP pioneered the use of Balanced Incomplete Block (BIB) spiraling, which made it possible to estimate group performance from relatively small samples of individual student responses (NCES, 2012).

NAEP has also been innovative in the assessment of content and skills not usually tested on a large scale, and in developing and using new forms of assessment to tap competencies not easily measured by paper-and-pencil tests. In recent years, for example, NAEP has tested the feasibility of assessing writing on computers in grade 4, and has developed an assessment of technology and engineering literacy, which was administered in 2014. All such studies have been highly visible because of NAEP's prominence. An assessment of collaborative problem solving would likely be equally high-profile.

NAEP's Influence on Policy and Practice

NAEP's importance and credibility has provided the assessment with a large influence on policy and practice at the state and local levels. Although no consequences are directly tied to NAEP results, schools, districts, and states have followed NAEP's lead and have revised standards and assessments in line with NAEP.

In some cases, the changes came about because of low NAEP scores. For example, California revised its reading curriculum after NAEP results showed the state was among the lowest performers of all states (DeVito & Koenig, 2000). In addition, one impetus for the development of the Common Core State Standards was the perceived discrepancies between the proportion of students deemed proficient on state tests and the proportion of students who reached the proficient level on NAEP (Rothman, 2011).

States also revised their standards and assessments to reflect those of NAEP. Connecticut modeled its reading frameworks on the NAEP reading framework, and the state of Washington did the same in mathematics. Officials from those states told a National Research Council committee that such changes would help ensure that NAEP results validate the state test results (DeVito & Koenig, 2000).

This level of influence suggests that NAEP should move with caution in developing new frameworks or in creating assessments in new domains. Such changes can reverberate far beyond the national assessment itself.

Collaborative Problem Solving: New Types of Data

An assessment of collaborative problem solving could in many ways represent a departure for NAEP. It could measure student competencies across a range of content areas, it could measure students' work in pairs or groups, and it could assess how students solve problems in addition to finding their solutions. Each of these raises questions about interpretation and inference.

Measuring a Cross-Content Framework

NAEP was created following the passage of the Elementary and Secondary Education Act of 1965, and it was designed to measure student progress in schools. Therefore, with the possible exception of the TEL assessment, all of NAEP's assessments have measured student performance on disciplines taught commonly in schools—mathematics, reading, science, U.S. history, writing, and so forth. The process for framework development, moreover, helps ensure that what is tested reflects a balance of what subject-matter experts and educators say is taught and what they believe should be taught.

A collaborative problem-solving assessment would differ from this practice. As noted in Chapter 3, there is no recognized framework for collaborative problem solving, as there is in the content areas. Moreover, such an assessment could cross disciplinary boundaries by measuring problem solving in multiple disciplines, such as language arts, mathematics, and science. It could also be discipline-neutral—the PISA test in problem solving measures student performance on problems that occur in real-world settings, such as buying a subway ticket.

These types of assessments might be difficult for educators to interpret. While teachers and administrators can readily see the instructional implications of results of, say, a science assessment, the implications of an assessment that crosses disciplines or is discipline-neutral are less clear. If students do relatively well, is that because they are good problem solvers in all disciplines, or do they do exceptionally well on science problems and not as well on mathematics problems? Or, if the test includes real-world discipline-neutral problems, what can schools do to improve student performance?

Furthermore, the assessment must be designed to isolate the problem-solving constructs from the underlying disciplinary content. If not, a student might perform poorly on a mathematics problem-solving task because of difficulties with mathematics, not problem solving.

The interpretive challenge is compounded by the fact that problem-solving abilities, particularly for real-world problems, can be developed outside of school. While student performance on many NAEP assessments can reflect both in-school and out-of-school learning—for example, students read outside of school, and the availability of books in the home is generally associated with higher reading performance—it is unclear to what extent schools can be responsible for developing the ability of students to solve problems they might not encounter in disciplinary subjects.

Regardless of whether the assessment includes discipline-based or discipline-neutral problems, some research in cognitive science suggests that problem solving is not a generic skill and is instead effective only within a domain. As a National Research Council report put it:

> Problem solvers confronted by a problem outside their area of expertise use…weak methods to try to constrain what would otherwise be very large search spaces when they are solving novel problems. In most situations, however, learners are expected to use strong methods—relatively specific algorithms which are particular to the domain that will make it possible to solve problems efficiently (Pellegrino & Hilton, 2012, p. 77).

Collaboration and Group Scores

Although NAEP does not provide scores for individual students, all of its assessments have been administered to individuals who take the tests, either on paper or on computer, on their own. In addition to the tests, students also complete a battery of contextual survey questions about themselves and about their instruction, which enables NAEP to draw associations between student characteristics and their performance.

An assessment of collaborative problem solving could change this. Although some versions of the assessment would measure the performance of a single student paired with a virtual agent, others would measure the performance of students in groups. Moreover, how well students collaborate with one another would form part of the score, in addition to their ability to solve a problem.

As Chapter 4 points out, there are two ways the assessment can be administered. One is to form pairs or groups of students, either in a single class or virtually, to collaborate on a problem. The second is for students to work with a virtual agent to solve a problem. Both raise issues of interpretation.

Some of the issues involved in the grouping approach should be familiar to any teacher who has had students perform group work. If students in the group come into the project with very different levels of ability, the higher-ability students are likely to perform much of the work. If that is the case, the difference in the students' efforts can be reflected in the "collaborative" portion of the score, but the "problem solving" portion will be more difficult to interpret.

Even if students come into a task with equal levels of ability, they may have different prior knowledge. For some students, the problem-solving task might be a situation they have encountered before, while for others, it could be completely unfamiliar. How can the assessment level the playing field for them?

The challenge is even more pronounced for English language learners. If one member of the group has limited English proficiency and the other(s) are proficient English speakers, how does that affect participation? How can accommodations be used so that they are fair to all?

Grouping also raises questions about associating contextual variables with student performance. Although NAEP does not provide scores for individual students, its design allows contextual variables to be matched with the characteristics of the students who fill out the surveys. But a collaborative problem-solving assessment would produce scores for a group's collective work. If more than one student participates in a group task and receives a group score, how can contextual variables be matched with students?

The use of agents can alleviate some of these concerns. In fact, the PISA framework concluded that the use of agents is the only fair way of assessing students' collaborative problem-solving competencies. As the framework document states:

> Group-level measurements are highly dependent on group composition and the individual skills of the participants (Kreijns, Kirschner, & Jochems, 2003; Rosen & Rimor, 2009). Fairly assigning a competency level to individuals working in a group where all group members can vary is impossible, because each individual's display of observable behaviour depends on the behaviour of the other group members....It has therefore been decided to place each individual student in collaborative problem-solving situations, where the team member(s) with whom the student has to collaborate is fully controlled. This is achieved by programming computer agents (PISA, 2013, p. 17).

However, while this solution might enable a fair assessment of students' problem-solving abilities by controlling extraneous variables, it creates an artificial environment that does not reflect how students might collaborate to solve problems in real-world settings in which the composition of groups will vary. Students might act differently with agents than they would with other people, and agents might not be programmed to react precisely the way other people might react to a student's actions. Thus the assessment should make clear the limitations of inferences from this approach.

Process Variables

In addition to measuring the quality of students' solutions to challenging problems, the collaborative problem-solving assessment can also measure students' steps along the way, that is, how they go about solving the problems. Computer based assessment makes possible the measurement of the quality of students' interactions with one another, whether and for how long they refer to reference documents, and whether they change their solution during the course of solving a problem, among other things.

Such information can be valuable. It can show, for example, whether particular strategies are associated with higher levels of performance, and whether particular strategies appear to be ineffective or counterproductive. This information makes the assessment important for teachers.

Indeed, NAEP has provided such information in past assessments. The 2011 writing assessment, the first to be administered on computers, reported 23 different actions students took as they read the writing prompts or drafted and revised their written responses. These included the use of a thesaurus; the use of spell-check; and the use of the cut, paste, and copy functions (NCES, 2012).

Yet while such information can be important, it also raises questions about interpreting problem-solving scores. Since the information can show what students do during the course of the assessment, it can show that they have, in fact, "learned" in the process of solving a problem, just as a video-game player can learn to avoid certain strategies when they run into roadblocks. In that case, how much do their problem-solving scores reflect what they know? Or is learning part of the competency of problem solving?

Conclusion

NAEP is a unique resource that has played a prominent and influential role as a respected and high-profile monitor of student achievement and as a pioneer in state-of-the-art assessment techniques. This dual role has led NAEP to exert a strong influence on policy and practice.

Because of NAEP's prominence, stakeholders need to exert caution in developing and implementing a new type of assessment. An assessment of collaborative problem solving would introduce a number of new elements to NAEP, which can pose challenges to interpretation. These include the use of multiple content areas (and possibly content-neutral problems), the use of student groups in test administrations, and the use of process variables. NCES and the Governing Board should be very clear about the information the assessment results provide, as well as its limitations.

Despite these challenges, an assessment of collaborative problem solving would represent a bold step that could strengthen NAEP's role as an assessment leader. As a so-called "21st century skill," collaborative problem solving is considered vital to students' future in the workplace. Policymakers and practitioners need to know how well students can demonstrate that skill, and NAEP is in an ideal position to provide that information.

References

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

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

American Management Association. (2010). AMA 2010 critical skills survey. Executive Summary. Retrieved from http://www.amanet.org/training/articles/3727.aspx.

American Psychological Association (2003). *How to build a better educational system: Jigsaw classrooms*. Retrieved from http://www.apa.org/research/action/jigsaw.aspx.

Arterberry, M., Cain, K., & Chopko, S. (2007). Collaborative problem solving in five-year-old children: Evidence of social facilitation and social loafing. *Educational Psychology*, 27, 577–596.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009. *Journal of Educational Data Mining, 1*, 3–17.

Bales, R. F. (1953). The equilibrium process in small groups. In T. Parson, R. F. Slater, & E. A. Shils (Eds.), Working papers in the theory of action (pp. 111-161). Glencoe, IL: Free Press.

Bales, R. F., & Strodtbeck, F. L. (1951). Phases in group problem-solving. Journal of Abnormal and Social Psychology, 46(4), 485-495.

Baranes, A. F., Oudeyer, P. Y., Gottlieb, J. (2014). The effects of task difficulty, novelty and the size of the search space on intrinsically motivated exploration. *Frontiers in Neuroscience, 8*, 317.

Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences, 12*, 307-359.

Black, E. B. (1955). A consideration of the rhetorical causes of breakdown in discussion. Communications Monographs, 22(1), 15-19.

Bowers, C. A., Jentsch, F., Salas, E., & Braun, C. C. (1998). Analyzing communication sequences for team training needs assessment, Human Factors, 40(4), 672 -679.

Bowers, C. A., Jentsch, F., & Salas, E. (2000). Establishing aircrew competencies: A comprehensive approach for identifying CRM training needs. In H. F. O'Neil & D. Andrews (Eds.). Aircrew training and assessment (pp. 67-84). Mahwah, NJ: Erlbaum.

Brannick, M. T., & Prince, C. (1997). An overview of team performance measurement. Team performance assessment and measurement. Theory, methods, and applications, 3.

Cannon-Bowers, J. A., Salas, E., & Converse, S. A. (1990) Cognitive psychology and team training: Training shared mental models and complex systems. Human Factors Society Bulletin, 33(12), 1–4.

Cannon-Bowers, J. A., & Salas, E. (1997). A framework for developing team performance measures in training. In M. T. Brannick, E. Salas, & C. W. Prince (Eds.), Team performance assessment and measurement: Theory, methods, and applications (pp. 45-62). Mahwah, NJ: Lawrence Erlbaum Associates.

Cannon-Bowers, J. A., Tannenbaum, S. I., Salas, E., & Volpe, C. E. (1995). Defining competencies and establishing team training requirements. In R. A. Guzzo & E. Salas (Eds.), Team effectiveness and decision making in organizations (pp. 333-381). San Francisco: Jossey-Bass.

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

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (pp. 7-75). Hillsdale, NJ: Erlbaum.

Cizek, G., & Bunch, M. B. (2007). *Standard setting*. New York: Sage.

Clark, H. (1996). *Using language*. Cambridge, United Kingdom: Cambridge University Press.

Cohen, E. G. (1982). Expectation states and interracial interaction in school settings. *American Review of Sociology, 8*, 209-235.

Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research, 64*, 1-36.

Cohen, E. G., Lotan, R., & Catanzarite, L. (1990). Treating status problems in the cooperative classroom. In S. Sharan (Ed.), *Cooperative learning: Theory and research* (203-230). New York, NY: Praeger.

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

Cooke, N. J., Gorman, J. C., & Winner, J. L. (2007). Team cognition. In R. S. Nickerson, S. Dumais, S. Lewandowsky, & T. J. Perfect (Eds.), *Handbook of Applied Cognition (2nd edition)* (239-270). West Sussex, England: John Wiley & Sons, Ltd.

Cooke, N. J., Salas, E., Cannon-Bowers, J. A., & Stout, R. J. (2000). Measuring team knowledge. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 42*, 151–173.

Cooke, N. J., Salas, E., Kiekel, P. A., & Bell, B. (2004). Advances in measuring team cognition. In E. Salas & S. M. Fiore (Eds.), *Team cognition: Understanding the factors that drive process and performance* (83–106). Washington, DC: American Psychological Association.

Dede, C. (2007). Reinventing the role of information and communications technologies in education. Yearbook of the National Society for the Study of Education, 106(2), 11-38.

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

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

DeVito, P. J., & Koenig, J. A., eds. (2000). *Reporting district-level NAEP data: Summary of a workshop*. Washington, DC: National Academies Press.

Dillenbourg, P., & Traum, D. (2006). Sharing solutions: Persistence and grounding in multi-modal collaborative problem solving. *The Journal of the Learning Sciences, 15,* 121-151.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109–132.

Entin, E. E., & Serfaty, D. (1999). Adaptive team coordination. Human Factors: The Journal of the Human Factors and Ergonomics Society, 41(2), 312-325.

Fadel, C., & Trilling, B. (2012). Twenty-first Century Skills and Competencies. In Encyclopedia of the Sciences of Learning (pp. 3353-3356). Springer US.

Fiore, S. M. (2008). Interdisciplinarity as teamwork: How the science of teams can inform team science. Small Group Research, 39(3), 251-277.

Fiore, S. M., Carter, D. R., & Asencio, R. (2015). Conflict, Trust, and Cohesion: Examining Affective and Attitudinal Factors in Science Teams. In E. Salas, W. B. Vessey, & A. X. Estrada (Eds.), Team Cohesion: Advances in Psychological Theory, Methods and Practice (pp. 271-301). Emerald Group Publishing Limited.

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, 52(2), 203-224.

Fiore, S. M. & Schooler, J. W. (2004). Process mapping and shared cognition: Teamwork and the development of shared problem models. In E. Salas & S. M. Fiore (Eds). Team Cognition: Understanding the factors that drive process and performance (pp. 133-152). Washington, DC: American Psychological Association.

Funke, J. (1991). Solving complex problems: Exploration and control of complex systems. In R. J. Sternberg & P. A. Frensch (Eds.), *Complex problem solving: Principles and mechanisms* (185-222). Hillsdale, NJ: Erlbaum.

Funke, J. (2010). Complex problem solving: a case for complex cognition? Cognitive Processing, 11(2), 133-142.

Gouran, D. S., & Hirokawa, R. Y. (1996). Functional theory and communication in decision making and problem-solving groups: An expanded view. In R. Y. Hirokawa & M. S. Poole (Eds.), Communication and group decision making (2nd ed., pp.55-80). Thousand Oaks, CA: Sage.

Graesser, A., & Foltz, P. (2013, June). *The PISA 2015 collaborative problem solving framework*. Paper presented at the 2013 Computer-Supported Collaborative Learning Conference. Madison, WI.

Graesser, A. C., Foltz, P. W., Rosen, Y., Shaffer, D. W., Forsyth, C., & Germany, M. (under review). Challenges of assessing collaborative problem solving. In E. Care, P. Griffin, and M. Wilson (Eds.), *Assessment and teaching of 21st century skills*. Heidelberg, Germany: Springer Publishers.

Graesser, A. C., Forsyth, C. M., & Foltz, P. (in press). Assessing conversation quality, reasoning, and problem solving performance with computer agents. In B. Csapo, J. Funke, and A. Schleicher (Eds.), *On the nature of problem solving: A look behind PISA 2012 problem solving assessment* (275-297). Heidelberg, Germany: OECD Series.

Greiff, S., Holt, D. V., & Funke, J. (2013). Perspectives on problem solving in educational assessment: analytical, interactive, and collaborative problem solving. Journal of Problem Solving, 5(2), 71-91.

Griffin, P. (2014). *Assessment and teaching of C21 Skills (ATC21S). Measuring collaborative skills: Challenges and opportunities*. Melbourne, Australia: University of Melbourne.

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

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

Griffin, P., Care, E., & Harding, S. (2015). Task characteristics and calibration. In P. Griffin and E. Care (Eds.) *Assessment and teaching of 21st century skills. Methods and approach: Educational assessment in an information age* (133-182). Dordrecht, Netherlands: Springer Science and Business Media.

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

Griffin, P., Care, E., & Wilson, M. (2016). *Measuring individual and group performance in collaborative problem solving*. Discovery Project DP160101678. University of Melbourne Australian Research Council.

Gurtner, A., Tschan, F., Semmer, N. K., & Nägele, C. (2007). Getting groups to develop good strategies: Effects of reflexivity interventions on team process, team performance, and shared mental models. Organizational Behavior and Human Decision Processes, 102(2), 127-142.

Halpin, P. F., & DeBoeck, P. (2013). Modelling dyadic interaction with Hawkes processes. Psychometrika, 78(4), 793-814.

Halpin, P. F., & von Davier, A. A. (2013). *Evaluating the roles of individual members in team interactions*. Presented at the National Council on Measurement in Education, San Francisco, CA.

Halpin P. F. & von Davier, A. A. (2017). Modeling collaboration using point processes. In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), Innovative Assessment of Collaboration. New York: Springer.

Halpin, P. F., von Davier, A. A., Hao, J., & Liu, L. (2014). *Modelling collaboration using point processes*. Presented at the Innovative Assessment of Collaboration, Washington DC.

Halpin, P. F., von Davier, A. A., Hao, J., & Liu, L. (2017). Measuring student engagement during collaboration. Journal of Educational Measurement, 54(1), 70-84.

Hao, J., Liu, L., von Davier, A., & Kyllonen, P., (2015). *Assessing collaborative problem solving with simulation based task*. Accepted paper for the 11th international conference on computer supported collaborative learning, Gothenburg, Sweden.

Hao, J., Liu, L., von Davier, A., Kyllonen, P., & Kitchen, C., (2016). *Collaborative problem solving skills versus collaboration outcomes: findings from statistical analysis and data mining*. Proceedings of the 9th International Conference on Educational Data Mining, Raleigh, NC.

Hao, J., von Davier, A. & Kyllonen, P. C. (in press). Towards a standardized assessment for collaborative problem solving (CPS): Practical challenges and strategies. In A. von Davier, M. Zhu, & P. C. Kyllonen (Eds). *Innovative assessment of collaboration*. Dordrecht, Netherlands: Springer.

Harnack, R. V. (1951). Competition and cooperation. Communication Studies, 3(1), 15-20.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77(1)*, 81-112.

He, Q., von Davier, M., Greiff, S., Steinhauer, E. W., & Borysewicz, P. B. (in press). Collaborative problem solving measures in the Programme for International Student Assessment (PISA). In A. von Davier, M. Zhu, & P. C. Kyllonen (Eds). *Innovative assessment of collaboration*. Dordrecht, Netherlands: Springer.

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin and E. Care (Eds.). Assessment and teaching of 21st century skills: Methods and approach. Educational Assessment in an Information Age. Dordrecht, Springer: 37-56

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research, 70(2)*, 151–179.

Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). The emerging conceptualization of groups as information processors. Psychological Bulletin, 121, 43-64.

Jackson, K. (2009). A little now for a lot later: A look at a Texas Advanced Placement Incentive Program. *The Journal of Human Resources, 45(3)*, 591-639.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1-73.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65*, 681–706.

Keltner, J. W. (1957). Group Discussion Processes. Longmans, Green.

Kerr, D. (in press). Using data mining results to improve educational video game design. *Journal of Educational Data Mining*.

Kerr, D., & Chung, G. K. W. K. (2012). Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Educational Data Mining, 4*, 144–182.

Kim, J. H., Gunn, D. V., Schuh, E., Phillips, B. C., Pagulayan, R. J., & Wixon, D. (2008). Tracking real-time user experience (TRUE): A comprehensive instrumentation solution for complex systems. *In Proceedings of the 26th annual SIGCHI Conference on Human Factors in Computing Systems*, 443–452.

Kyllonen, P. C. (2017). Socio-emotional and self-management variables in learning and assessment. In A. A. Rupp & J. P. Leighton (Eds.), *The Handbook of cognition and assessment: Frameworks, methodologies, and applications (pp. 174-197)*. Hoboken, NJ: Wiley-Blackwell.

Kyllonen, P. C., Zhu, M., & von Davier, A. A. (2017). Introduction: Innovative Assessment of Collaboration. In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), Innovative Assessment of Collaboration. Switzerland: Springer Nature.

Lai, E. R. (2011). *Collaboration: A literature review*. Retrieved from http://images.pearsonassessments.com/images/tmrs/Collaboration-Review.pdf.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Alstyne, M. V. (2009). Computational Social Science. Science, 323(5915), 721-723.

Letsky, M., Warner, N., Fiore, S. M., & Smith, C. (Eds.). (2008). Macrocognition in Teams: Theories and Methodologies. London: Ashgate Publishers.

Levy, F., & Murnane, R. J. (2013). Trends in Routine and Non-routine Tasks in Occupations, United States, 1960 to 2009 [Figure]. *Dancing with robots: Human skills for computerized work*. Retrieved from http://content.thirdway.org/publications/714/Dancing-With-Robots.pdf.

Levy, R. (2014). *Dynamic Bayesian network modeling of game based diagnostic assessments* (CRESST Report 8037). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Littman, M. L. (2000, October). Review: computer language games. In International Conference on Computers and Games (pp. 396-404). Springer Berlin Heidelberg.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher, 41(9)*, 352-362.

Liu, L., von Davier, A., Hao, J., Kyllonen, P., & Zapata-Rivera, D. (2015). A tough nut to crack: Measuring collaborative problem solving. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds). *Handbook of research on computational tools for real-world skill development* (344–359). Hershey, PA: IGI-Global.

Lowenstein G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin, 116(1)*, 75–98.

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science, 4*, 333-370.

Mazzeo, J., Lazer, S., & Zieky, M. (2006). Monitoring educational progress with group-score assessments. In R. L. Brennan (Ed.), *Educational Measurement, 4th edition*. (17-64; 681-699). Westport, CT: Praeger.

Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology, 85(2)*, 273-283.

Mazur, E. (1997). *Peer instruction. A user's manual.* Upper Saddle River, NJ: Prentice-Hall.

McGrath, J. E. (1984). *Groups: Interaction and performance*. Englewood Cliffs, NJ: Prentice Hall.

Melbourne Graduate School of Education, University of Melbourne (Australia). *Assessment and teaching of 21st century skills project (ATC21S)*. Retrieved from http://ATC21S.org/index.php/resources/white-papers/.

Mislevy, R. J., Almond, R. G., Dibello, L. V., Jenkins, F., Steinberg, L. S., Yan, D., & Senturk, D. (2002). *Modeling conditional probabilities in complex educational assessments* (CSE Technical Report 580). Los Angeles, CA: The National Center for Research on Evaluation, Standards, Student Testing, University of California, Los Angeles.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives, 1, 3-67.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., & Lukas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. M. Williamson, I. I., Bejar, & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (15–48). Mahwah, NJ: Erlbaum.

Mohammed, S., & Dumville, B. C. (2001). Team mental models in a team knowledge framework: Expanding theory and measurement across disciplinary boundaries. *Journal of Organizational Behavior 22*, 89-106.

Morgan, B., Keshtkar, F., Duan, Y., & Graesser, A. C. (2012). Using state transition networks to analyze multi-party conversations in a serious game. In. S. A. Cerri, & B. Clancey (Eds.), *Proceedings of the 11th International Conference on Intelligent Tutoring Systems (ITS 2012)* (162-167). Berlin: Springer-Verlag.

Morgan, B., Keshtkar, F., Graesser, A., & Shaffer, D. W. (2013). Automating the mentor in a serious game: A discourse analysis using finite state machines. In C. Stephanidis (Ed.), *Human Computer Interaction International 2013*, 374 (591–595). Berlin: Springer.

Mosier, K. L., & Fischer, U. M. (2010). Judgment and decision making by individuals and teams: Issues, models, and applications. Reviews of Human Factors and Ergonomics, 6(1), 198-256.

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

National Assessment Governing Board (NAGB) (2012). *NAEP 12th grade preparedness research*. Retrieved from https://www.nagb.org/what-we-do/preparedness-research/types-of-research/jss.html.

National Center for Education Statistics (NCES) (2012). Writing 2011: *The National Assessment of Educational Progress at grades 8 and 12*. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf.

National Center for Education Statistics (2012). The Nation's Report Card: Hands-On Tasks (HOT) and Interactive Computer Tests (ICT) for the 2009 Science Assessment. Washington, DC: Author.

National Center for Education Statistics (NCES) (2012). Improving the *measurement of socioeconomic status for the National Assessment of Educational Progress: A theoretical foundation*. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/researchcenter/Socioeconomic_Factors.pdf.

National Center for Education Statistics (2012). *Leading assessment into the future*. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/Future_of_NAEP_Panel_White_Paper.pdf.

National Center for Education Statistics (2014). *The Nation's Report Card: A first look: 2013 mathematics and reading* (NCES 2014-451). Retrieved from https://nces.ed.gov/nationsreportcard/subject/publications/main2013/pdf/2014451.pdf.

National Center for Education Statistics (2015). The Nation's Report Card: 2015 Mathematics and Reading Assessments. Publication Number NCES 2015136. Washington, DC: Author.

National Center for Education Statistics (2016). *The Nation's Report Card: 2014 Technology and Engineering Literacy*. Washington, DC: Author.

National Research Council (2011). *Assessing 21st century skills*. Washington, DC: National Academies Press.

Nelson, B. (2007). Exploring the use of individualized, reflective guidance in an educational multi-user virtual environment. *Journal of Science Education and Technology 16(1)*, 83–97.

Newell, A., & Simon, H. A. (1972). Human Problem Solving. Englewood Cliffs, NJ: Prentice-Hall.

Ogle, L. T., & Dabbs, P. A. (1998). The media's mixed record in reporting test results. In G. I. Maeroff (Ed.), *Imaging education: The media and schools in America* (85-100). New York: Teachers College Press.

O'Neil, H. F., Jr., Chung, G. K. W. K., & Brown, R. (1997). Use of networked simulations as a context to measure team competencies. In H. F. O'Neil, Jr. (Ed.), Workforce readiness: Competencies and assessment (pp. 411-452). Mahwah, NJ: Lawrence Erlbaum Associates.

O'Neil, H. F., Jr., & Chuang, S. H., & Chung, G. K. W. K. (2004). *Issues in the computer-based assessment of collaborative problem solving*. Center for the Study of Evaluation (CSE) Report 620. Retrieved from http://www.cse.ucla.edu/products/reports/r620.pdf.

O'Neil, H. F., Jr., & Chuang, S. H. (2008). Measuring collaborative problem solving in low-stakes tests. In E. L. Baker, J. Dickieson, W. Wulfeck, & H. F. O'Neil (Eds.), *Assessment of problem solving using simulations* (177–199). Mahwah: Lawrence Erlbaum Associates.

Orasanu, J. (1994). Shared problem models and flight crew performance. In N. Johnston, N. McDonald and R. Fuller (Eds.), Aviation psychology in practice (pp. 255-285). Brookfield, VT: Ashgate.

Orasanu, J. (2005). Crew collaboration in space: A naturalistic decision making perspective. Aviation, Space, and Environmental Medicine, 76(Supplement 1), B154-B163.

Orasanu, J., & Salas, E. (1993). Team decision-making in complex environments. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), Decision-making in action: Models and methods (pp. 158 -171). Norwood: Ablex

Organisation for Economic Co-operation and Development (OECD) (2010). *PISA 2012 field trial problem solving framework*. Retrieved from www.oecd.org/dataoe cd/8/42/46962005.pdf.

Organisation for Economic Co-operation and Development (OECD) (2013). *PISA 2015 collaborative problem solving frameworks*. Retrieved from http://www.oecd.org/pisa/pisaproducts/pisa2015draftframeworks.htm.

Organisation for Economic Co-operation and Development (OECD) (2013). *PISA 2012 results: Ready to learn: Students' engagement, drive and self-beliefs (Volume III)*. Retrieved from http://www.oecd.org/pisa/keyfindings/pisa-2012-results-volume-iii.htm.

Parker, G. M. (1994/2003) Cross-functional teams. San Francisco, CA: Jossey-Bass.

Pellegrino, J. W., & Hilton, M. L. (Eds.) (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: National Academies Press.

Polya, G. (1971). *How to solve it*. Princeton, NJ: Princeton University Press.

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

Romero, C., & Ventura, S. (2005). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33(1)*, 135–146.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (2010). *Handbook of educational data mining*. Cleveland, OH: CRC Press.

Rosé, C. P., Howley, I., Wen, M., Yang, D., & Ferschke, O. (2017). Assessment of Discussion in Learning Contexts. In von Davier, Alina A., Zhu, Mengxiao, Kyllonen, Patrick C. (Eds.) Innovative Assessment of Collaboration (pp. 81-94). Springer International Publishing.

Rosé, C., Wang, Y. C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning, 3*, 237–271.

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, 380-406.

Rosen, Y., & Rimor, R. (2012). Teaching and assessing problem solving in online collaborative environment. Teacher education programs and online learning tools: Innovations in teacher preparation, 82-97. Hershey, PA: Information Science Reference.

Rothman, R. (2011). *Something in common: The Common Core State Standards and the next chapter in American education*. Cambridge, MA: Harvard Education Press.

Russell, S. J., & Norvig, P. (2003), *Artificial intelligence: A modern approach* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Salas, E., Cooke, N. J., & Rosen, M. A. (2008). On teams, teamwork, and team performance: Discoveries and developments. Human Factors, 50(3), 540–547.

Salas, E., Sims, D. E., & Burke, C. S. (2005). Is there a "big 5" in teamwork? Small Group Research, 36(5), 555-599.

Salen, K., & Zimmerman, E. (2004). *Rules of play: Game design fundamentals*. Cambridge, MA: MIT Press.

Schippers, M. C., Den Hartog, D. N., Koopman, P. L., & Wienk, J. A. (2003). Diversity and team outcomes: The moderating effects of outcome interdependence and group longevity and the mediating effect of reflexivity. Journal of Organizational Behavior, 24, 779-802.

Schmitz, M. J., & Winskel, H. (2008). Towards effective partnerships in a collaborative problem-solving task. British Journal of Educational Psychology, 78(4), 581-596.

Scoular, C. (2016). *Towards a template for collaborative problem solving tasks*. Unpublished doctoral dissertation, University of Melbourne, Australia.

Shachar, H., & Sharan, S. (1994). Talking, relating and achieving: Effects of cooperative learning and whole-class instruction. *Cognition and Instruction, 12(4)*, 313-353.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2007). *An assessment for learning system called ACED: Designing for learning effectiveness and accessibility* (Research Rep. No. RR-07-27). Princeton, NJ: Educational Testing Service.

Soller, A., & Stevens, R. (2008). Applications of stochastic analyses for collaborative learning and cognitive assessment. In G. R. Hancock & K. M. Samuelson (Eds.), *Advances in latent variable mixture models*, Charlotte, N.C.: Information Age Publishing, Inc.

Stecher, B. M., & Hamilton, L. S. (2014). Measuring Hard-to-Measure Student Competencies: A Research and Development Plan. Research Report. RAND Corporation. PO Box 2138, Santa Monica, CA 90407-2138.

Stokols, D., Misra, S. Moser, R., Hall, K. L., & Taylor, B. (2008). The ecology of team science: Understanding contextual influences on transdisciplinary collaboration. American Journal of Preventive Medicine, 35, 2, S96-S115.

Thissen, D., & Wainer, H. (Eds.) (2001). *Test scoring*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thompson, E. J. (1967). *Organizations in action*. New York, NY: McGraw-Hill.

Trilling, B., & Fadel, C. (2009). 21st century skills: Learning for life in our times. John Wiley & Sons.

Tuckman, B. W. (1965). Developmental sequence in small groups. Psychological Bulletin, 63(6), 384-399.

University of Reading (2014, June). *Turing Test success marks milestone in computing history*. http://www.reading.ac.uk/news-and-events/releases/PR583836.aspx.

United States Department of Labor. Secretary's Commission on Achieving Necessary Skills. (1991). What work requires of schools: a SCANS report for America 2000. Secretary's Commission on Achieving Necessary Skills, US Department of Labor.

United States Department of Labor. (2000). *Skills and tasks for jobs: A scans report for America (1-5)*. Retrieved from http://wdr.doleta.gov/research/FullText_Documents/1999_35.pdf.

VanLehn, K., & Martin, J. (1998). Evaluation of an assessment system based on Bayesian student modeling. *International Journal of Artificial Intelligence in Education, 8*, 179–221.

Vomlel, J. (2004). Bayesian networks in educational testing. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 12(1 supp)*, 83–100.

Von Davier, A., & Halpin, P., (2013). *Collaborative problem solving and the assessment of cognitive skills: psychometric considerations*. Research Report ETS RR-13-41. Princeton, NJ: Educational Testing Service.

Von Davier, A. A., Zhu, M., & Kyllonen, P. (2017). *Innovative assessment of collaboration*. Springer International Publishing.

Wagner, T. (2008). Rigor redefined. Educational Leadership, 66(2), 20-24.

Wagner, T. (2010). *Creating Innovators: The Making of Young People Who Will Change the World*. New York, NY: Scribner.

Wang, Z., Hao, J., Liu, L., Chen, L., & von Davier, A., (2015). *Automated classification of collaborative problem solving skills based on chat messages in a science simulation task*. Paper presented at InGroup 10th Annual Conference, Pittsburgh, PA.

Webb, N. M. (1991). Task-related verbal interaction and mathematical learning in small groups. *Research in Mathematics Education, 22(5)*, 366–389.

Webb, N. M. (1995). Group collaboration in assessment: Multiple objectives, processes, and outcomes. *Educational Evaluation and Policy Analysis, 17(2)*, 239–261.

Webb, N. M., Nemer, K. M., Chizhik, A. W., & Sugrue, B. (1998). Equity issues in collaborative group assessment: Group composition and performance. *American Educational Research Journal, 35(4)*, 607–651.

West, M. A. (2000). Reflexivity, revolution and innovation in work teams. In M. M. Beyerlein, D. A. Johnson, & S. T. Beyerlein (Eds.), Product development teams (Vol. 5, pp. 1–29). Stamford, CT: JAI Press.

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

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching. Special Issue: Learning Progressions. 46 (6)*, 716–730.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science, 330(6004)*, 686–688.

World Economic Forum. (2015). *New Vision for Education: Unlocking the Potential of Technology*. Retrieved from http://www3.weforum.org/docs/WEFUSA_NewVisionforEducation_Report2015.pdf.

Wright, B., & Masters, G. (1983) *Rating scale analysis*. Chicago: MESA Press.