Educational Research and Innovation

The Nature of Problem Solving

USING RESEARCH TO INSPIRE 21ST CENTURY LEARNING

Edited by Benő Csapó and Joachim Funke

Centre for Educational Research and Innovation

OECD

Educational Research and Innovation

THE NATURE OF PROBLEM SOLVING

USING RESEARCH TO INSPIRE 21ST CENTURY LEARNING

Edited by Benő Csapó and Joachim Funke

OECD
BETTER POLICIES FOR BETTER LIVES

This work is published on the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the OECD member countries.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Please cite this publication as:
Csapó, B. and J. Funke (eds.) (2017), The Nature of Problem Solving: Using Research to Inspire 21st Century Learning, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264273955-en

ISSN: 2076-9660 (print)
ISSN: 2076-9679 (online)

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

Photo credits:
Fotolia/alphaspirit

Foreword

The demands on learners and thus education systems are evolving fast. In the past, education was about teaching people something. Now, it's about making sure that students develop a reliable compass and the navigation skills to find their own way through an increasingly uncertain, volatile and ambiguous world. These days, we no longer know exactly how things will unfold, often we are surprised and need to learn from the extraordinary, and sometimes we make mistakes along the way. And it will often be the mistakes and failures, when properly understood, that create the context for learning and growth. A generation ago, teachers could expect that what they taught would last a lifetime for their students. Today, teachers need to prepare students for more rapid economic and social change than ever before, for jobs that have not yet been created, to use technologies that have not yet been invented, and to solve social problems that we don't yet know will arise.

The dilemma for educators is that the kind of skills that are easiest to teach and easiest to test, are also the skills that are easiest to digitise, automate and outsource. There is no question that state-of-the-art disciplinary knowledge will always remain necessary. Innovative or creative people generally have specialised skills in a field of knowledge or a practice. And as much as "learning to learn" skills are important, we always learn by learning something. However, success in life and work is no longer mainly about reproducing content knowledge, but about extrapolating from what we know and applying that knowledge in novel situations. Put simply, the world no longer rewards people just for what they know – Google knows everything – but for what they can do with what they know. Problem solving is at the heart of this, the capacity of an individual to engage in cognitive processing to understand and resolve problem situations where a method of solution is not immediately obvious.

Conventionally our approach to problems in schooling is to break them down into manageable pieces, and then to teach students the techniques to solve them. But today individuals create value by synthesising disparate parts. This is about curiosity, open-mindedness, making connections between ideas that previously seemed unrelated, which requires being familiar with and receptive to knowledge in other fields than our own. If we spend our whole life in a silo of a single discipline, we will not gain the imaginative skills to connect the dots, which is where the next invention will come from.

Perhaps most importantly, in today's schools, students typically learn individually and at the end of the school year, we certify their individual achievements. But the more interdependent the world becomes, the more we rely on great collaborators and orchestrators who are able to join others to collaboratively solve problems in life, work and citizenship. Innovation, too, is now rarely the product of individuals working in isolation but an outcome of how we mobilise, share and link knowledge. So schools now need to prepare students for a world in which many people need to collaborate with people of diverse cultural origins, and appreciate different ideas, perspectives and values; a world in which people need to decide how to trust and collaborate across such differences; and a world in which their lives will be affected by issues that transcend national boundaries. Expressed differently, schools need to drive a shift from a world where knowledge is stacked up somewhere depreciating rapidly in value towards a world in which the enriching power of collaborative problem-solving activities is increasing.

These shifts in the demand for knowledge and skills are well understood and documented, and to some extent they are even intuitive. Not least, many school curricula highlight the importance of individual and social problem-solving skills. And yet, surprisingly little is known about the extent to which education systems deliver on these skills. This is not just because school subjects continue to be shaped by traditional disciplinary contexts. It is also because educators have few reliable metrics to observe the problem-solving skills of their students - and what doesn't get assessed doesn't get done.

The OECD Programme for International Student Assessment (PISA) sought to address this. Its 2012 assessment contained the first international metric of individual problem-solving skills and the 2015 assessment took this further, assessing collaborative problem-solving skills. The results turned out to be extremely interesting, in part because they showed that strong problem-solving skills are not an automatic by product of strong disciplinary knowledge and skills. For example, Korea and Japan, which both did very well on the PISA mathematics test, came out even stronger on the PISA assessment of problem-solving skills. In contrast, top mathematics performer Shanghai did relatively less well in problem solving. Such results suggest that it is worth educators devoting more attention to how problem-solving skills are developed both in disciplinary and cross-disciplinary contexts.

But while problem solving is a fairly intuitive and all-pervasive concept, what has been missing so far is a strong conceptual and methodological basis for the definition, operationalisation and measurement of such skills. This book fills that gap. It explores the structure of the problem-solving domain, examines the conceptual underpinning of the PISA assessment of problem solving and studies empirical results. Equally important, it lays out methodological avenues for a deeper analysis of the assessment results, including the study of specific problem-solving strategies through log-file data.

In doing so, the book provides experts and practitioners with the tools to better understand the nature of problem-solving skills but also with a foundation to translate advanced analyses into new pedagogies to foster better problem-solving skills.

Andreas Schleicher

Andreas Schleicher

Director for Education and Skills
Special Advisor to the Secretary-General

Acknowledgements

Thanks to all the people who contributed to this book and helped to finalise a process that took much longer than expected! Special thanks to Julia Karl (Heidelberg) who helped us preparing a standardised and printable version of all manuscripts. And thanks to Marion Lammarsch (Heidelberg) for help with converting text from LaTex to Word. Thanks to the OECD staff, in particular Francesco Avvisati, Sophie Limoges and Rachel Linden, for their valuable support during the production process, and to Sally Hinchcliffe for the thoughtful language editing of the manuscript.

This book stems mainly from the collaboration of the members of the OECD Problem Solving Expert Group (PEG) of PISA 2012. This PEG group started its works in 2009 and finished their official work in 2014. The group consisted of the following eight members:

- Joachim Funke (Chair), Heidelberg University, Germany
- Benő Csapó, University of Szeged, Hungary (ex officio PGB representative)
- John Dossey, Illinois State University, United States
- Art Graesser, University of Memphis, United States
- Detlev Leutner, Duisburg-Essen University, Germany
- Richard Mayer, University of California, United States
- Tan Ming Ming, Ministry of Education, Singapore
- Romain Martin, University of Luxembourg, Luxembourg

Members of the PEG group have already been involved with problem solving for a long time, and their meetings under the umbrella of the PISA work inspired a number of other meetings and co-operative studies involving people from other organisations and institutions. Influenced by the creative atmosphere of the PEG meetings, some of the members met and presented their work together at other professional meetings as well. Amongst these were the annual meetings at Szeged University in the framework of Szeged Workshop of Educational Evaluation (SWEE; since 2009 a yearly repeated event), the TAO days in Luxemburg (March, 2011), the AERA meeting in New Orleans (April, 2011), two symposia at the EARLI biennial meeting in Exeter (September, 2011), the European Conference of Psychology in Istanbul (July, 2011), two symposia at the International Conference on Psychology in Cape Town (July, 2012), and more recently the “Celebrating Problem Solving” conference at the University of Szeged (November 2015). Many related journal articles have been published in the meantime – too many to be listed here.

These productive meetings brought together researchers interested in problem solving, assessment of cognitive skills, and technology based assessment, and so initiated empirical works in the overlapping areas of these special fields. For example, as already mentioned, one of the major shifts from PISA 2003 problem solving to PISA 2012 problem solving was the shift from paper-and-pencil based to computer-based assessment that required strong interactions between item developers and the group taking care of the technical implementation. Within a rather short time scale, software tools had to be developed and implemented that allow for the necessities in international large-scale assessment studies (e.g. preparing for more than 100 different languages, different character sets including left-to-right and right-to-left, and different levels of computer equipment).

The PEG group was supported by a wonderful team from ACER (Australian Council for Educational Research, Melbourne, Australia): the "trio" consisting of Barry McCrae, Ray Philpot, and Dara Ramalingam. They prepared meetings and materials in a fantastic way and helped us through a jungle of dates, deadlines, and data. Ray Adams worked as Interim Chair in the beginning of the project. All of this contributed to the success of PISA 2012.

Maybe unique in the history of PISA expert groups, this community of researchers, while developing the assessment framework and creating the instruments discovered the potentials of an emerging field: the possibilities offered by computerised, dynamic, interactive assessment of problem solving. Using multimedia and simulation to present the test tasks, capturing students' responses in novel ways, logging student's activities and using log-file analyses for exploring cognitive processes, perseverance and motivation have opened new and exciting directions for research. They have been continuing their collaboration far beyond their task in the 2012 assessment cycle.

The individual chapters in this book have been reviewed by members of the group and by reviewers from the OECD. This process hopefully helped to improve the quality of the chapters. At the same time, these activities delayed the publication process a bit.

Lastly we would like to extend our thanks to Andreas Schleicher for his support throughout the publication process.

Table of contents

Figures

Tables

Executive summary

Problem solving is one of the key competencies humans need in a world full of changes, uncertainty and surprise. It is needed in all those situations where we have no routine response at hand. Problem solving requires the intelligent exploration of the world around us, it requires strategies for efficient knowledge acquisition about unknown situations, and it requires creative application of the knowledge available or that can be gathered during the process. The world is full of problems because we strive for so many ambitious goals – but the world is also full of solutions because of the extraordinary competencies of humans who search for and find them.

PISA 2012 addressed the issue of problem-solving competency with two firsts: 1) the use of computer-based assessment methods; and 2) the use of interactive problem solving in its framework (as distinct from static problem solving which was the focus of former PISA problem solving assessments in 2003). Both new approaches require a rethinking of theories, concepts and tools. This reflection will be done within this book. It explains why we made the shift from static to interactive problem solving.

This book presents the background and the leading ideas behind the development of the PISA 2012 assessment of problem solving, as well as some results from research collaborations that originated within the group of experts who guided the development of this assessment.

Part I gives an overview of problem solving and its assessment. In Chapter 1, Benő Csapó and Joachim Funke highlight the relevance of problem solving within a world that relies less and less on routine behaviour and increasingly requires non-routine, problem-solving behaviour. They underline the need for innovative assessments to understand how improved school practices contribute to the development of problem-solving skills in learners. In Chapter 2, Jens Fleischer and colleagues use data from PISA 2003 to demonstrate the importance of analytical problem solving as a cross-curricular competency. Starting with the observation that German students' analytical problem solving skills are above the OECD average but their maths skills are only average, they develop the cognitive potential exploitation hypothesis: good problem-solving skills could be harnessed to develop mathematics skills. In Chapter 3, Magda Osman argues that complex problems can be represented as decision-making tasks under conditions of uncertainty. This shift in emphasis offers a better description of the skills that underpin effective problem solving, with a focus on the subjective judgments people make about the controllability of the problem-solving context. In Chapter 4, John Dossey describes the mathematical perspective and points to the strong relation between mathematics and problem solving. He emphasises the important role of metacognition and self-regulation in both solving real problems and in the practice of mathematics.

Part II introduces dynamic problem solving as a new perspective within PISA 2012. In Chapter 5, Dara Ramalingam and colleagues present the PISA 2012 definition of problem solving, outline the development of the assessment framework and discuss its key organising elements, presenting examples of static and interactive problems from this domain. In Chapter 6, Samuel Greiff and Joachim Funke describe the core concept of interactive problem solving in PISA 2012 as the interplay of knowledge acquisition and knowledge application to reach a given goal state. Interactive problem solving differs from static problem solving because some of the information needed to solve the problem has to be found during interaction processes. In Chapter 7, Andreas Fischer and colleagues describe typical human strategies and shortcomings in coping with complex problems, summarise some of the most influential theories on cognitive aspects of complex problem solving, and present experimental and psychometric research that led to a shift in focus from static to dynamic problem solving.

Part III deals with data from different studies presented here because they influenced the thinking around the PISA 2012 problem-solving assessment. In Chapter 8, Gyöngyvér Molnár and colleagues report results from a Hungarian study of students from primary and secondary schools who worked with static and interactive problem scenarios. They found that students' ability to solve these dynamic problems appears to increase with age and grade level, thus providing evidence that education can influence the development of these skills. At the same time, the measured construct is psychometrically sound and stable between cohorts. In Chapter 9, Ray Philpot and colleagues describe analyses of item characteristics based on the responses of students to PISA 2012 problem solving items. They identified four factors that made problems more or less difficult, which could be helpful for test developers but also researchers and educators interested in developing learners from novice to expert problem-solvers in particular domains. In Chapter 10, Philipp Sonnleitner and colleagues discuss the challenges and opportunities of computer-based complex problem solving in the classroom through the example of Genetics Lab, a newly developed and psychometrically sound computer-based microworld that emphasises usability and acceptance amongst students.

Part IV presents ideas arising from the new computer-based presentation format used in PISA 2012. In Chapter 11, Nathan Zoanetti and Patrick Griffin explore the use of the data from log files, which were not available from paper-and-pencil tests and offer additional insights into students' procedures and strategies during their work on a problem. They allow researchers to assess not just the final result of problem solving but also the problem-solving process. In Chapter 12, Krisztina Tóth and colleauges show the long road from log-file data to knowledge about individuals' problem solving activities. They present examples of the usefulness of clustering and visualising process data. In Chapter 13, David Tobinski and Annemarie Fritz present a new assessment tool EcoSphere, which is a simulation framework for testing and training human behavior in complex systems. Its speciality is the explicit assessment of previously acquired content knowledge.

Part V deals with future issues. Three years after PISA 2012 focused on individual problem-solving competencies, PISA 2015 moved into the new and innovative domain of collaborative problem solving, defined as "the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution." In Chapter 14, Esther Care and Patrick Griffin present an alternative assessment of collaborative problem solving that inspired the PISA assessment while being clearly distinct from it. They deal with human-to-human collaboration, in contrast to the human-to-computer design eventually used. In Chapter 15, Arthur Graesser and colleagues, explore the use of "conversational agents" (intelligent computer-based systems that interact with a human) and how dialogues and even trialogues with two agents can be used for new types of assessment.

In summary, these chapters should help to better understand the PISA 2012 results published by OECD in 2014, by giving background information about the framework and its specific realisation. They should also help to better understand the advent of assessment of collaborative problem solving in PISA 2015 – an issue that reflects the increasing importance of collaboration beyond the need for individual problem solving. We hope that this book fulfills its purpose not just of providing information about the background, but of shaping the future of problem-solving research and assessment.

Benő Csapó, Joachim Funke, Andreas Schleicher

PART I

Problem solving: Overview of the domain

Chapter 1

The development and assessment of problem solving in 21st-century schools

By
Benő Csapó
MTA-SZTE Research Group on the Development of Competencies, University of Szeged, Hungary.
and Joachim Funke
Department of Psychology, Heidelberg University, Germany.

The skills considered most essential in our modern societies are often called 21st-century skills. Problem solving is clearly one of them. Students will be expected to work in novel environments, face problems they have never seen and apply domain-general reasoning skills that are not tied to specific contents. Computerised dynamic problem solving can be used to create just such an interactive problem situation in order to assess these skills. It may therefore form the basis for a type of assessment which helps answer the question of how well schools are preparing their students for an unknown future. This chapter shows how education systems may benefit from such an assessment. It reviews educational methods that have aimed at developing students' higher-order thinking skills and indicates how experiences with these approaches can be used to improve problem solving, from direct teaching, through content-based methods, to innovative classroom processes. It outlines the evolution of large-scale assessment programmes, shows how assessing problem solving adds value and, finally, identifies some directions for further research.

The preparation of this article was supported during the years 2007 to 2014 by grants Fu 173/11–1, Fu 173/14–1 and Fu 173/14–2 issued by the German Research Foundation (DFG) to the second author within the "Competence Models for Assessing Individual Learning Outcomes and Evaluating Educational Processes" Priority Programme (SPP 1293).

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

Introduction

Several features characterise developed societies at the beginning of the 21st century, but two of these have prevailing consequences for schooling: 1) the determining role of knowledge and skills possessed by their citizens; and 2) the rapid changes apparent in all areas of life. Students therfore not only have to master more knowledge to be able to live a successful life in these societies, but they also have to master a different type of knowledge which enables them to succeed in a rapidly changing environment. This means that if schools only prepare their students for current expectations, their knowledge and skills will be outdated by the time they have to use them in their private life and in the world of work.

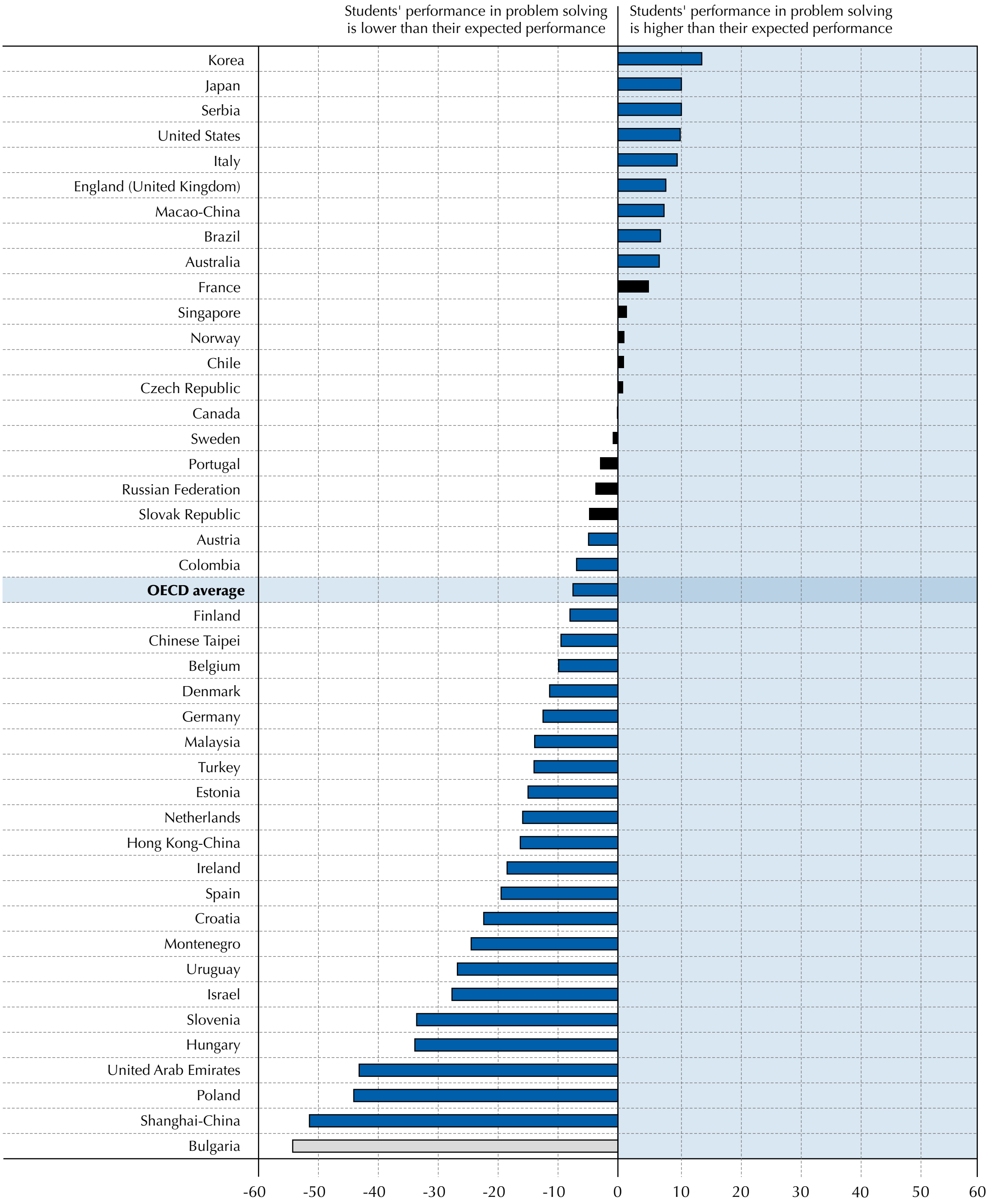
Reflecting these expectations, a new term has emerged in recent decades in the discussion about the goals of education: a family of skills considered the most essential in modern societies, often called 21st-century skills. Lists of these skills usually include creativity, innovation, communication, collaboration, decision making, social skills, cross-cultural skills, information and communications technology (ICT) literacy, civic literacy, media literacy, critical thinking, learning to learn and problem solving (see for example Binkley et al., 2012). However, when these skills are examined in detail, it often turns out that there are very few research results related to some of them, and some of them are difficult to define and specify precisely enough to be measured. Furthermore, in most cases, it is even more difficult to find methods for developing them.

Problem solving stands out in this group as it has been researched for several decades (Frensch and Funke, 1995; Funke and Frensch, 2007; Klieme, 2004; Mayer, 1992). Although problem solving is a general term and the corresponding field of research is very broad, several forms of problem solving are well defined and there are clear distinctions between some of them, for example between domain-specific and domain-general as well as between analytic and complex problem solving (Fischer, Greiff and Funke, 2012; Greiff, Holt and Funke, 2013; Greiff et al., 2013; Wüstenberg, Greiff and Funke, 2012). It is also known from research that complex problem solving, like most complex skills, develops over a long period (Molnár, Greiff and Csapó, 2013; Wüstenberg et al., 2014).

Because of its complex nature and long developmental time, there is little firm evidence showing how different educational methods stimulate its enhancement. The long-term impact of education on other cognitive abilities has been studied since the very beginning of research on intelligence. The general conclusion of such studies is that education improves general cognitive abilities: there is a clear relationship between years spent at school and level of cognitive abilities. A recent longitudinal study indicated that students taking academic tracks made greater gains than those receiving vocational education (Becker et al., 2012); furthermore, the results of a long-term longitudinal study suggest that the impact of education may be subject-specific rather than general (Ritchie, Bates and Deary, 2015). It was the problem solving assessment in the 2012 Programme for International Student Assessment (PISA) which first provided data that demonstrated that there are large differences between educational systems in terms of improving a well-defined cognitive ability.

As problem solving was the fourth (innovative) domain of the 2012 assessment, it was possible to compare the impacts of different education systems on performance in three main domains and on problem solving. The results of these comparisons have shown that even countries which are very good at teaching reading, mathematics and science literacy may be weaker in developing students' problem-solving abilities (OECD, 2014). Figure 1.1 demonstrates these differences. This analysis applied a regression model to predict the problem-solving performance of the countries participating in this assessment, based on their performance in the three main domains. The figure shows the difference between expected levels of problem solving (predicted by the regression analysis) and the measured levels.

Figure 1.1 **Relative performance in problem solving**



Source: OECD (2014), PISA 2012 Results: Creative Problem Solving (Volume V)
Student Skills in Tackling REal-Life Problems
StatLink http://dx.doi.org/10.1787/9789264208070-en, p. 69, Fig. V.2.15.

There are some countries where mathematics, science, and reading performance is high or has improved a great deal in the past decade, for example Shanghai-China and Poland, but problem-solving performance is relatively low. Other countries, such as Korea and Japan, perform well in the main domains and even better in problem solving than could be predicted simply on the basis of their other results. Students from the United States also perform better than expected (Dossey and Funke, 2016). In a country-by-country comparison of performance, a variety of patterns can be identified, clearly indicating that improving problem solving requires something other than teaching the main domains well. In this assessment, about 32% of the variance of problem-solving skills is not explained by the reading, mathematics and science scores. Although the reasons for these differences cannot be precisely identified based on the available data from this assessment, it is clear that the teaching and learning methods used in some countries are more effective at developing problem solving than others.

These results may provide an impetus for research on methods of improving problem solving in an educational context. Although this may be an emerging field of research, the intention of developing cognitive abilities and research in this field is not new at all. Experiences from previous attempts at improving cognitive abilities may also be used in the field of problem solving. In past work, there has been a close interaction between teaching and assessment processes both in research and in educational practice. It is therefore reasonable to also deal with these two aspects of problem solving in parallel. Bloom’s classical taxonomies were used both to describe and operationalise educational objectives on the one hand, and as the foundation of assessment instruments on the other (Bloom et al., 1956). In modern educational systems, curricula and assessment frameworks are often developed in parallel. Making cognitive constructs measurable is a precondition for systematic improvement; conversely, the emergence of new teaching methods calls for the development of specific assessment instruments.

Following this logic, the next sections describe the evolution of teaching methods which aim at improving the quality of knowledge and developing higher-order thinking skills, and show how they may serve as models for improving problem solving as well. They then outline the changes in large-scale international assessment, in parallel with those observed in teaching, that enable them to measure newer attributes of knowledge, and summarise the role dynamic problem solving plays in these developments. Finally, the chapter outlines the prospects for assessing and developing problem solving as a result of the PISA assessment and the research it has initiated.

Educational methods aimed at improving the quality of knowledge

Traditional conceptions of learning focused on memorising facts and figures and reproducing them in an unchanged context. Teachers were considered the sources of knowledge, and classroom work was typically teacher-directed. On the other hand, the intentions of cultivating students’ minds, preparing them to reason better, and helping them to be able to use their knowledge beyond the school context are as old as organised schooling itself. Although these were the declared goals, they often remained slogans simply because of the lack of scientific background and practical knowledge, and so rote learning dominated everyday practice. Collaboration, creativity and problem solving were also listed among educational goals in the past century, but in reality only a small proportion of students were actually required to apply these skills in their work and in everyday life.

A number of instructional experiments in the 20th century produced remarkable results beyond rote learning, but they failed to spread to entire education systems. As rapid technological, social and economic development at the beginning of this century has made it essential for most citizens to acquire new cognitive skills at a high level in order to find a job in the knowledge economy, these earlier innovations may need to be re-evaluated. This section considers four kinds of earlier or more recent developments which may be relevant to problem solving in education. Although, as the previous section noted, it is not known why some education systems are more successful at developing problem-solving skills than others, the intensity and frequency with which they use these approaches may contribute to their results.

Direct teaching of thinking skills

The rise of research on general cognitive abilities has inspired a number of educational methods aimed at improving these abilities. The first developmental programmes were independent of any school subject and aimed at developing very general cognitive constructs or even intelligence. Depending on the interpretation of general cognitive abilities, it is still disputed which components can be taught and to what extent, if at all. In educational contexts, models of plastic general ability have proved to be more fruitful than the rigid single-factor conception of intelligence. This shift is also supported by recent research in cognitive neuroscience (for an elaboration of this issue, see Adey et al., 2007).

One of the most well-known programmes in this category was Feuerstein's Instrumental Enrichment Program, which was effective in some applications, for example children with special educational needs and socially handicapped students, but did not produce the expected results when more broadly applied (see Blagg, 1991). Klauer devised a series of better-specified intervention programmes for the development of inductive reasoning (based on an operational definition), starting with materials for the direct training of young and handicapped children (Klauer, 1989) and continuing with more broadly applicable training programmes, including the development of problem solving (Klauer and Phye, 1994).

The main difficulty with these direct, independent, stand-alone programmes was that they required specific training (extra efforts beyond the usual instructional processes) which meant they were usually short. Their lasting impact and the transfer of any effects to other skills or broader contexts often remained insufficient. On the other hand, these intervention studies resulted in a number of particular intervention methods and provided rich experiences of aspects of trainability such as the sensitive age, effect size, transfer and duration of effects, which could then be used in designing instructional programmes. Experiments on the direct development of problem solving (such as Kretzschmar and Süß, 2015) may have a similar positive influence on educational applications.

Content-based methods: Integrating the teaching of disciplinary content and improving reasoning

A more fruitful approach seems to be the use of improved teaching materials for developing specific reasoning skills, as the disciplinary materials are there anyway and the time students spend mastering a subject can also be used to improve general cognitive abilities . These approaches are often called infusion methods, embedding or enrichment; each term indicates that some extra developmental effect is added to the traditional instructional processes (see Csapó, 1999 for a general overview of this approach).

Cognitive Acceleration through Science Education (CASE) is one of the best elaborated and most broadly tested developmental programmes in this category (Adey and Shayer, 1994; Adey, Shayer and Yates, 2001). Its embedded training effects not only improve general cognitive abilities, but also aid in a better mastery of science content. It thus meets the requirements of other progressive instructional approaches, such as "teaching for understanding" and "teaching for transfer".

Other content-based methods for developing problem solving use slightly different approaches aimed at developing problem-solving competencies. Problem-based instruction (PBI, see Mergendoller, Maxwell and Bellisimo, 2006 for an example) or problem-based learning (PBL, Hmelo-Silver, 2004; for a meta-analysis of its effects, see Dochy et al., 2003) organise teaching and learning around larger, complex, natural problems which often require active reasoning as well as the mobilisation and integration of knowledge from several traditional school subjects.

These integrated, content-based methods are often used to increase students' interest and motivation. As they are usually implemented within instructional time, they have the potential for broader application. The results of training experiments indicate that durable impacts can only be expected from long-term interventions, and long-term intervention can only realistically be implemented if they use improved (enriched) regular teaching processes.

Enhancing instruction to improve problem-solving abilities

A number of specific improvements in instructional processes which offer the possibility of improving problem solving have also been studied. Implementing them systematically in everyday school processes would have a measurable cumulative impact.

Using representations may contribute to a better understanding of learning materials in general, but representing complex problems properly is the first step in a successful problem solving process. Training students to use representations, especially multiple representations (e.g. depictive and descriptive), and practising the transformations between representations may facilitate the development of problem solving as well (Schnotz et al., 2010). In a similar way, visualisation (for which computerised teaching materials offer excellent opportunities) supports the mastery of content knowledge and at the same time improves problem-solving skills (Wang et al., 2013).

Computer-based simulations of problem situations are an important condition for the assessment of dynamic problem solving, and may be used for developmental purposes as well (e.g. Rivers and Vockell, 2006). Similar to the idea of simulation is gamification or game-based learning. This rapidly developing area of educational technology engages students in well-structured serious games which require them to practise certain reasoning skills. Simulation may help students understand some complicated phenomena, while game-based learning combines learning and entertainment and so makes learning more exciting and improves student perseverance (Connolly et al., 2012; Hwang and Wu, 2012). These methods have been tested for the development of problem solving in a number of different areas and contexts (for example Chang et al., 2012; Rowe et al., 2011; Spires et al., 2011).

A number of studies have shown that training students to better monitor their own learning processes, thus facilitating self-regulated learning and metacognition, may also contribute to the improvement of problem solving (Montague, 2008; Perels, Gürtler and Schmitz, 2005).

Schoenfeld has generalised his work on teaching methods of mathematical problem solving and developed his theory of goal-oriented decision making, which is then applicable in broader teaching contexts, including the improvement of problem solving (Schoenfeld, 2011). He builds on monitoring and self-regulation, which are components of several other innovative methods as well. Metacognition and self-regulated learning help students to develop their own problem-oriented study processes.

Global approaches to improving interest, motivation and the quality of learning

Modern constructivist theories describe learning as an interaction with the environment, with the teacher's role being to provide students with a stimulating physical and social environment and to guide their students through their own developmental processes (e.g. Dumont, Istance and Benavides, 2010). A number of methods to enhance environmental effects have recently been piloted and introduced into educational practice; these are sometimes called "innovative learning environments" (OECD, 2013a) or, if enhanced with ICT, "powerful learning environments"

(De Corte, 1990; De Corte et al., 2003; Gerjets and Hesse, 2004).

Some of these methods have already been introduced into everyday school practices, such as working in pairs and several forms of group work which build on the benefits of social interaction. Collaborative learning, especially its computer-mediated form, also creates excellent opportunities for the development of problem solving (Uribe, Klein and Sullivan, 2003). These types of activities already form a bridge towards collaborative problem solving (e.g. Rummel and Spada, 2005), which was the innovative domain of the PISA 2015 assessment.

Discovery learning builds on students' curiosity and may be very motivating (Alfieri et al., 2011). Similarly, inquiry-based learning takes place through students' active observation and experimentation; inquiry-based science education (IBSE) has recently become especially popular (Furtak et al., 2012). IBSE aims to introduce the basic processes of scientific research into the practices of school instruction. One of its declared goals is to develop scientific reasoning and other higher-order thinking skills by creating and testing hypotheses. This is one of the most broadly studied methods used to revitalise science education; indeed, the European Union has supported more than 20 IBSE projects during the past decade.

Developing the scope of international assessment programmes

Countries' educational cultures differ with respect to the depth and frequency of the application of these approaches to teaching. It seems a plausible hypothesis that the methods described above better prepare students for an unknown future than traditional direct teaching. To test this hypothesis and to measure the cumulative impact of these innovative methods on real learning outcomes, however, also requires the development of innovative assessment instruments. This need initiated the assessment of problem solving in PISA 2003 (OECD, 2005). Nine years later, the development of technology made it possible to overcome the limitations of paper-based assessments for assessing problem-solving skills at the level of education systems. This section provides an overview of how large-scale assessment projects have evolved from the early curriculum-based surveys through the application-oriented conception of literacy to the assessment of higher-order thinking skills.

This overview considers a three-dimensional (curricular content knowledge, application of knowledge and psychological dimension) approach to the goals of learning (Csapó, 2010) and shows how they complement each other.

Curriculum-based content-focused assessments

The first international assessment programmes were the International Association for the Evaluation of Educational Achievement (IEA) studies in the 1970s and 1980s. They were based on an analysis of the curricula of participating countries. Bloom's taxonomies (Bloom et al., 1956) governed the analysis of curricula and the development of assessment frameworks. The content assessed was thus almost identical with what was taught at schools.

However, one of the most obvious experiences of school education is that students usually cannot apply the knowledge they have mastered at school in a new context. In other words, the transfer of knowledge is not automatic. Some traditional forms of teaching result in inert knowledge, as they are not effective in facilitating the application and improving the transfer of knowledge.

These experiences initiated changes in the IEA surveys, and the Trends in International Mathematics and Science Study (TIMSS) assessments, which have been carried out regularly every four years since 1995, measure a broader range of competencies. The most recent framework of TIMSS mathematics and science assessment deals with broad categories such as knowing, applying and reasoning (Mullis and Martin, 2013).

Assessing the application of knowledge

Although assessing the ability to apply knowledge was a goal of early international assessment programmes (Bloom's taxonomy also considers application), they did not elaborate the scope and areas of transfer and application played a secondary role. The cognitive revolution in psychology in the decades preceding the launch of the PISA assessments initiated new directions for research on human information processing as well as for knowledge and skills as outcomes of learning.

These new insights and empirical research results have found their way into educational applications. The OECD"s PISA programme has also drawn from this knowledge base. The OECD's Definition and Selection of Key Competencies (DeSeCo) programme drew on results from the cognitive sciences and interpreted the conception of competencies in the context of school education (Rychen and Salganik, 2001). The first PISA frameworks reinterpreted the conception of literacy and elaborated the definitions of reading literacy, mathematical literacy and scientific literacy as forms of knowledge that are applicable in typical contexts of modern societies.

In sum, the early assessment programmes deduced their frameworks from the knowledge of particular scientific disciplines (science and mathematics), while PISA added a new dimension: the application of knowledge. Literacies – in other words, the conception of broadly applicable knowledge and skills (e.g. PISA assessment frameworks) – can be deduced from the social environments, expectations and demands of modern life. More precisely, it is not what students have been taught, but what they are expected to know that directs the development of frameworks, beyond taking into account the most recent results from cognitive research and research on learning and instruction in general (OECD, 2013b).

Assessing general cognitive skills: The psychological dimensions of knowledge

Neither the knowledge base of mathematics and science and other disciplines, nor the present social demands for applicable knowledge provides a perfect source to determine the skills that will be needed in the future. It seems plausible that students can be prepared for an unknown future by developing their general cognitive skills.

This new need to assess a further dimension of knowledge is marked by the large number of publications dealing with this new future orientation. Some deal with the assessment of higher-order thinking skills (Schraw and Robinson, 2011), 21st-century skills (Griffin and Care, 2015; Mayrath et al., 2012), a variety of competencies (Hartig, Klieme and Leutner, 2008), and, especially, problem solving (Baker et al., 2008; Ifenthaler et al., 2011).

Dynamic problem solving, the innovative domain of PISA 2012, has all the features of such an assessment, evaluating how well students are prepared to solve problems where they do not have ready-made, well-practised routine solutions. Nevertheless, to integrate such a domain into the system of assessments and to develop a framework for it, really requires an anchoring in the psychology of human reasoning: knowledge of human information processing in general, and the psychology of problem solving in particular.

The relevance of assessing problem solving for PISA

When students are expected to work in a novel environment and face problems they have never seen, they cannot use content knowledge they have mastered before but they can apply their domain-general reasoning skills that are not tied to specific content. They have to understand the situation, create a model, and, through interaction with a particular segment of the environment, explore how it responds and behaves. The knowledge gained from this interaction can then be used to build a model, generalise new knowledge and subsequently use it to solve the actual problem.

Computerised dynamic problem solving creates such a problem situation. It may thus form the basis for a type of assessment which helps us to answer the question of how well schools prepare their students for an unknown future.

In a dynamic problem-solving task, students face a system that simulates a real system and behaves similarly to it. The problem-solving process typically has two main phases (see Funke, 2001): a knowledge acquisition phase and a knowledge application phase. Problem solvers have to mobilise a number of operational reasoning skills as well as higher-order thinking skills in both phases. First, they have to understand the simulated system as a whole and then identify the relevant variables of the system. In this phase, they may apply their analogical thinking in search of already known analogous systems. Then they have to manipulate the variables systematically and observe how changing the value of one variable affects the value of other variables. In the meantime, they may use a number of combinatorial and classification operations, propositional reasoning skills, etc. By organising the results of their observations, they induce some rules and test if the rules are valid. When they are convinced that the rules they have just discovered are correct, they begin to apply this newly mastered knowledge. In the knowledge application phase, they may again use a number of further reasoning skills (Greiff and Funke, 2009; Greiff, Wüstenberg and Funke, 2012).

Knowledge acquisition is an important process in learning the content of school subjects, and knowledge application is required for PISA literacy tasks when knowledge mastered in one situation is applied to another one. However, dynamic problem solving assesses students' cognitive skills in a clearer, better-controlled situation, where the results are less masked by the availability of the knowledge students have mastered before.

Conclusions for further research and development

To illustrate the nature of the difficulties we face when we consider the prospects for fostering problem solving in schools, we may use an analogy from physics. We know that nuclear fusion is possible and that it produces immense energy. It is derived from theories of physics, and we know that the sun produces its energy via nuclear fusion. Humans have also created conditions in which nuclear fusion takes place: the hydrogen bomb, which also generates an incredible amount of energy. Thus, it is proven that nuclear fusion works, its mechanisms are known, but we do not have the technology yet that could harness nuclear fusion to produce energy for everyday purposes.

Similarly, even if it is proven that general cognitive abilities are amenable, and the skills underpinning problem solving have already been successfully developed in a number of well-controlled experiments, there is still a long road before we have school curricula and teaching and learning practices that develop problem solving much better than today's schools. Looking back at the history of research on problem solving, we see that the majority of studies have been exploring the mechanisms of problem solving, mostly under laboratory conditions and with simple static tasks. Tests that are usable in an educational context are a recent development, although the availability of good measurement instruments is a precondition for the design of intervention experiments and controlled implementation of large-scale developmental programmes. The PISA 2012 assessment is a ground-breaking enterprise from this perspective, as it demonstrates the possibilities of computer-based assessment in a number of different educational systems.

For now, researchers should first outline a roadmap of research that would produce evidence for the improvement of practice. Such a map would probably include the study of school practices in countries which perform better in problem solving than expected on the basis of their achievement in the main domains (see Buchwald, Fleischer and Leutner, 2015, who discuss a "cognitive potential exploitation hypothesis"). Although it is very difficult to import good practices from one complex educational culture into another, a deeper understanding of how schooling helps students to become better problem solvers in one country may help to improve practices in other countries as well.

Studying the structure and development of problem solving in an educational context also still offers a great deal of potential. Cross-sectional assessments with a large number of background variables may produce results in a short period. In particular, understanding the role of precursors and the early development of component skills may be beneficial, as recent research has underscored the importance of early development. Longitudinal studies, on the other hand, may take a longer time but provide data on real developmental trajectories. Studying the factors that determine development should include a broad range of affective variables.

Finally, experiments for developing problem solving in real educational contexts should be encouraged. Taking into account the multi-causal character of such development, it will be impossible to find a single, simple solution. Experiences from former programmes for developing other higher-order thinking skills suggest that stable outcomes may only result from continuous stimulation of cognition over a long period. As there is no reasonable way to teach problem solving in separate courses, the most promising solutions are those that embed training in problem solving into regular school activities. A number of intervention points should be explored, including standards, assessment frameworks, textbooks, classroom practices and extra-curricular activities. Technology may again be helpful in this matter: simulation and gamification, especially online educational games, may offer unexplored potential for developing problem solving.

References

Adey, P., B. Csapó, A. Demetriou, J. Hautamäki, and M. Shayer (2007), "Can we be intelligent about intelligence? Why education needs the concept of plastic general ability", *Educational Research Review*, Vol. 2/2, pp. 75-97, http://dx.doi.org/10.1016/j.edurev.2007.05.001.

Adey, P. and M. Shayer (1994), *Really Raising Standards: Cognitive Intervention and Academic Achievement*, Routledge, London.

Adey, P., M. Shayer and C. Yates (2001), *Thinking Science: The Curriculum Materials of the CASE Project*, 3rd edition, Nelson Thornes, London.

Alfieri, L., P.J. Brooks, N.J. Aldrich and H.R. Tenenbaum (2011), "Does discovery-based instruction enhance learning?", *Journal of Educational Psychology*, Vol. 103/1, pp. 1-18.

Baker, E., J. Dickieson, W. Wulfeck and H.F. O'Neil (eds.) (2008), *Assessment of Problem Solving Using Simulations*, Erlbaum, New York.

Becker, M., O. Lüdtke, U. Trautwein, O. Kölle and J. Baumert (2012), "The differential effects of school tracking on psychometric intelligence: Do academic-track schools make students smarter?", *Journal of Educational Psychology*, Vol. 104/3, pp. 682-699, http://dx.doi.org/10.1037/a0027608.

Binkley, M. et al. (2012), "Defining twenty-first century skills", in P. Griffin, B. McGaw and E. Care (eds.), *Assessment and Teaching of 21st Century Skills,* Springer, Dordrecht, pp. 17-66.

Blagg, N. (1991), *Can we Teach Intelligence? A Comprehensive Evaluation of Feuerstein's Instrumental Enrichment Program*, Erlbaum, Hillsdale, NJ.

Bloom, B.S. et al. (1956), *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*, David McKay Company, New York.

Buchwald, F., J. Fleischer and D. Leutner (2015), "A field experimental study of analytical problem solving competence – Investigating effects of training and transfer", *Thinking Skills and Creativity*, Vol. 18, pp. 18-31, http://dx.doi.org/10.1016/j.tsc.2015.04.009.

Chang, K.E., L.J. Wu, S.E. Weng and Y.T. Sung (2012), "Embedding game-based problem-solving phase into problem-posing system for mathematics learning", *Computers & Education*, Vol. 58/2, pp. 775-786, http://dx.doi.org/10.1016/j.compedu.2011.10.002.

Connolly, T.M. et al. (2012), "A systematic literature review of empirical evidence on computer games and serious games", *Computers & Education*, Vol. 59/2, pp. 661-686, http://dx.doi.org/10.1016/j.compedu.2012.03.004.

Csapó, B. (2010), "Goals of learning and the organization of knowledge", *Zeitschrift für Pädagogik*, Vol. 56 (Beiheft), pp. 12-27.

Csapó, B. (1999), "Improving thinking through the content of teaching", in J.H.M. Hamers, J.E.H. van Luit and B. Csapó (eds.), *Teaching and Learning Thinking Skills,* Swets and Zeitlinger, Lisse, pp. 37-62.

De Corte, E. (1990), "Towards powerful learning environments for the acquisition of problem-solving skills", *European Journal of Psychology of Education*, Vol. 5/1, pp. 5-19.

De Corte, E., L. Verschaffel, N.J. Entwistle and J. van Merriënboer (eds.) (2003), *Powerful Learning Environments: Unravelling Basic Components and Dimensions,* Elsevier, Oxford.

Dochy, F., M. Segers, P. Van den Bossche and D. Gijbels (2003), "Effects of problem-based learning: A meta-analysis", *Learning and Instruction*, Vol. 13/5, pp. 533-568, http://dx.doi.org/10.1016/S0959-4752(02)00025-7.

Dossey, J.A. and J. Funke (2016), "Canadian and United States students' performances on the OECD's PISA 2012 problem-solving assessment", *Canadian Journal of Science, Mathematics and Technology Education*, Vol. 16, pp. 92-108, http://dx.doi.org/10.1080/14926156.2015.1119332.

Dumont, H., D. Istance and F. Benavides (eds.) (2010), *The Nature of Learning: Using Research to Inspire Practice*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264086487-en.

Feuerstein, R., Y. Rand, M. Hoffman and R. Miller (1980), *Instrumental Enrichment: An Intervention Program for Cognitive Modifiability*, University Park Press, Baltimore.

Fischer, A., S. Greiff and J. Funke (2012). "The process of solving complex problems", *Journal of Problem Solving*, Vol.4/1, pp. 19-42, http://dx.doi.org/10.7771/1932-6246.1118.

Frensch, P.A. and J. Funke (eds.) (1995), *Complex Problem Solving: The European Perspective*, Erlbaum, Hillsdale, NJ.

Funke, J. (2010), "Complex problem solving: A case for complex cognition?", *Cognitive Processing*, Vol. 11/2, pp. 133-142, http://dx.doi.org/10.1007/s10339-009-0345-0.

Funke, J. (2001), "Dynamic systems as tools for analysing human judgement", *Thinking & Reasoning*, Vol. 7/1, pp. 69-89, http://dx.doi.org/10.1080/13546780042000046.

Funke, J. and P.A. Frensch (2007), "Complex problem solving: The European perspective – 10 years after", in D.H. Jonassen (ed.), *Learning to Solve Complex Scientific Problems*, Erlbaum, New York, pp. 25-47.

Furtak, E.M., T. Seidel, H. Iverson and D.C. Briggs (2012), "Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis", *Review of Educational Research*, Vol. 82/3, pp. 300-329, http://dx.doi.org/10.3102/0034654312457206.

Gerjets, P. and F. Hesse (2004), "When are powerful learning environments effective? The role of learner activities and of students' conceptions of educational technology", *International Journal of Educational Research*, Vol. 41/6, pp. 445-465, http://dx.doi.org/10.1016/j.ijer.2005.08.011.

Greiff, S. and J. Funke (2009), "Measuring complex problem solving: The MicroDYN approach", in F. Scheuermann and J. Björnsson (eds.), *The Transition to Computer-Based Assessment: New Approaches to Skills Assessment and Implications for Large-Scale Testing*, European Commission, Ispra, Italy, pp. 157-163.

Greiff, S., D.V. Holt and J. Funke (2013), "Perspectives on problem solving in educational assessment: Analytical, interactive, and collaborative problem solving", *Journal of Problem Solving*, Vol. 5/2, pp. 71-91, http://dx.doi.org/10.7771/1932-6246.1153

Greiff, S., S. Wüstenberg and J. Funke (2012), "Dynamic problem solving: A new assessment perspective", *Applied Psychological Measurement*, Vol. 36/3, pp. 189-213, http://dx.doi.org/10.1177/0146621612439620.

Greiff, S., S. Wüstenberg, G. Molnár, A. Fischer, J. Funke and B. Csapó (2013), "Complex problem solving in educational settings – Something beyond *g*: Concept, assessment, measurement invariance, and construct validity, *Journal of Educational Psychology*, Vol. 105/2, pp. 364-379, http://dx.doi.org/10.1037/a0031856.

Griffin, P. and E. Care (eds.) (2015), *Assessment and Teaching of 21st Century Skills: Methods and Approach*, Springer, Heidelberg, http://dx.doi.org/10.1007/978-94-017-9395-7.

Hartig, J., E. Klieme and D. Leutner (eds.) (2008), *Assessment of Competencies in Educational Contexts*, Hogrefe, Göttingen.

Hmelo-Silver, C.E. (2004), "Problem-based learning: What and how do students learn?", *Educational Psychology Review*, Vol. 16/3, pp. 235-266, http://dx.doi.org/10.1023/B:EDPR.0000034022.16470.f3.

Hwang, G.J. and P.H. Wu (2012), "Advancements and trends in digital game-based learning research: A review of publications in selected journals from 2001 to 2010", *British Journal of Educational Technology*, Vol. 43/1, pp. 6-10, http://dx.doi.org/10.1111/j.1467-8535.2011.01242.x.

Ifenthaler, D., et al. (eds.) (2011), *Multiple Perspectives on Problem Solving and Learning in the Digital Age*, Springer, New York.

Klauer, K.J. (1989), *Denktraining für Kinder*, Hogrefe, Göttingen.

Klauer, K.J. and G. Phye (1994), *Cognitive Training for Children. A Developmental Program of Inductive Reasoning and Problem Solving*, Hogrefe, Seattle.

Klieme, E. (2004), "Assessment of cross-curricular problem-solving competencies", in J.H. Moskowitz and M. Stephens (eds.), *Comparing Learning Outcomes: International Assessments and Education Policy* (pp. 81–107). London: Routledge.

Kretzschmar, A. and H.-M. Süß (2015), "A study on the training of complex problem solving competence", *Journal of Dynamic Decision Making*, Vol.1/1, pp. 1-15, http://dx.doi.org/10.11588/jddm.2015.1.15455.

Mayer, R.E. (1992), *Thinking, Problem Solving, Cognition*, 2nd edition, WH Freeman, New York.

Mayrath, M.C., J. Clarke-Midura, D.H. Robinson and G. Schraw (eds.) (2012), *Technology-Based Assessments for 21st Century Skills: Theoretical and Practical Implications from Modern Research*, Information Age Publishing, Charlotte, NC.

Mergendoller, J.R., N.L. Maxwell and Y. Bellisimo (2006), "The effectiveness of problem-based instruction: A comparative study of instructional methods and student characteristics", *Interdisciplinary Journal of Problem-Based Learning*, Vol. 1/2, pp. 1-22, http://dx.doi.org/10.7771/1541-5015.1026.

Molnár, G., S. Greiff and B. Csapó (2013), "Inductive reasoning, domain specific and complex problem solving: Relations and development", *Thinking Skills and Creativity*, Vol. 9, pp. 35-45, http://dx.doi.org/10.1016/j.tsc.2013.03.002.

Montague, M. (2008), "Self-regulation strategies to improve mathematical problem solving for students with learning disabilities", *Learning Disability Quarterly*, Vol. 31/1, pp. 37-44, http://dx.doi.org/10.2307/30035524.

Mullis, I.V.S. and M.O. Martin (eds.) (2013), *TIMSS 2015 Assessment Frameworks*, TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.

OECD (2014), *PISA 2012 Results: Creative Problem Solving (Volume V): Students' Skills in Tackling Real-Life Problems*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264208070-en.

OECD (2013a), *Innovative Learning Environments*, Centre for Educational Research and Innovation, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264203488-en.

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

OECD (2005), *Problem Solving for Tomorrow's World: First Measures of Cross-Curricular Competencies from PISA 2003*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264006430-en.

Perels, F., T. Gürtler and B. Schmitz (2005), "Training of self-regulatory and problem-solving competence", *Learning and Instruction*, Vol. 15/2, pp. 123-139, http://dx.doi.org/10.1016/j.learninstruc.2005.04.010.

Ritchie, S.J., T.C. Bates and I.J. Deary (2015), "Is education associated with improvements in general cognitive ability, or in specific skills?", *Developmental Psychology*, Vol. 51/5, pp. 573-582, http://dx.doi.org/10.1037/a0038981.

Rivers, R. H. and E. Vockell (2006), "Computer simulations to stimulate scientific problem solving", *Journal of Research in Science Teaching*, Vol. 24/5, pp. 403-415, http://dx.doi.org/10.1002/tea.3660240504.

Rowe, J.P., L.R. Shores, B.W. Mott and J.C. Lester (2011), "Integrating learning, problem solving, and engagement in narrative-centered learning environments", *International Journal of Artificial Intelligence in Education*, Vol. 21/1-2, pp. 115-133, http://dx.doi.org/10.3233/JAI-2011-019.

Rummel, N. and H. Spada (2005), "Learning to collaborate: An instructional approach to promoting collaborative problem solving in computer-mediated settings", *Journal of the Learning Sciences*, Vol. 14/2, pp. 201-241.

Rychen, D.S. and L.H. Salganik (eds.) (2001), *Defining and Selecting Key Competencies*, Hogrefe & Huber, Göttingen.

Schnotz, W., C. Baadte, A. Müller and R. Rasch (2010), "Creative thinking and problem solving with depictive and descriptive representations", in L. Verschaffel, E. De Corte T. de Jong and J. Elen (eds.), *Use of Representations in Reasoning and Problem Solving,* Routledge, New York, pp. 11-35.

Schoenfeld, A.H. (2011), *How We Think: A Theory of Goal-Oriented Decision Making and its Educational Applications*, Routledge, London.

Schraw, G. and D.H. Robinson (eds.) (2011), *Assessment of Higher Order Thinking Skills*, Information Age Publishing, Charlotte, NC.

Spires, H.A., J.P. Rowe, B.W. Mott and J.C. Lester (2011), "Problem solving and game-based learning: Effects of middle grade students' hypothesis testing strategies on learning outcomes", *Journal of Educational Computing Research*, Vol. 44/4, pp. 453-472, http://dx.doi.org/10.2190/EC.44.4.e.

Uribe, D., J.D. Klein and H. Sullivan (2003), "The effect of computer-mediated collaborative learning on solving ill-defined problems", *Educational Technology Research and Development*, Vol. 51/1, pp. 5-19.

Wang, M., B. Wu, N.-S. Chen and J.M Spector (2013), "Connecting problem-solving and knowledge-construction processes in a visualization-based learning environment", *Computers & Education*, Vol. 68, pp. 293-306, http://dx.doi.org/10.1016/j.compedu.2013.05.004.

Wüstenberg, S., S. Greiff and J. Funke (2012), "Complex problem solving – More than reasoning?", *Intelligence*, Vol. 40/1, pp. 1-14, http://dx.doi.org/10.1016/j.intell.2011.11.003.

Wüstenberg, S., S. Greiff, G. Molnár and J. Funke (2014), "Cross-national gender differences in complex problem solving and their determinants", *Learning and Individual Differences*, Vol. 29/1, pp. 18-29, http://dx.doi.org/10.1016/j.lindif.2013.10.006.

Chapter 2

Analytical problem solving: Potentials and manifestations

By

Jens Fleischer, Florian Buchwald, Detlev Leutner

Department of Instructional Psychology, University of Duisburg-Essen, Germany.

Joachim Wirth

Department of Research on Learning and Instruction, Ruhr-University Bochum, Germany.

and Stefan Rumann

Institute of Chemistry Education, University of Duisburg-Essen, Germany.

Being able to solve problems is a crucially important outcome of almost all domains of education, as well as a requirement for future learning and acheivement. This chapter summarises the results from the cross-curricular analytical problem solving element of the 2003 Programme for International Student Assessment (PISA). The results from Germany indicated that schools may not be sufficiently exploiting students' cognitive potential to develop their subject-specific competencies. To investigate this hypothesis and find ways to make use of this potential, the chapter describes research on the cognitive components and the structure of analytical problem-solving competence and its relation to mathematical competence. It concludes with results from two experimental studies aiming to foster students' mathematical competence by training in analytical problem-solving competence.

The preparation of this article was supported by grant LE 645/12–2 and LE 645/12–3 from the German Research Foundation (DFG) in the Priority Programme "Competence Models for Assessing Individual Learning Outcomes and Evaluating Educational Processes" (SPP 1293).

Introduction

The competence to solve problems is of crucial importance in almost all domains of learning and achievement, both in and out of school. Accordingly, educational standards in various subject areas define it as a subject-specific competence (for example AAAS, 1993; NCTM, 2000; Blum et al., 2006). However, problem solving is also seen as a cross-curricular competence that is acquired and applied in the school context and that is vital for success in the world of work (for example OECD, 2005a). Problem solving is also an object of interest to various research approaches. Educational researchers have developed methods to teach problem-solving skills in diverse learning situations in various school subjects (for example Charles and Silver, 1988; Bodner and Herron, 2002; Törner et al., 2007). In psychology, especially European cognitive psychology, research on problem solving has a long tradition. Its roots lie in the experimental research conducted in Gestalt psychology and the psychology of thinking in the first half of the 20th century (for example Duncker, 1935; Wertheimer, 1945).

Despite differing categorisations of problems and problem solving, there is consensus that a problem consists of a problem situation (initial state), a more or less well-defined goal state, and a solution method that is not immediately apparent to the problem solver (Mayer, 1992). Thus, problem solving requires logically deriving and processing certain information in order to successfully solve the problem. In contrast to a simple task, a problem is a non-routine situation for which no standard solution methods are readily at hand (Mayer and Wittrock, 2006). To use Dörner's (1976) terminology, a barrier between the initial state and the desired goal state makes it unclear how to transform the initial state into the goal state. The solution of problems can thus be defined as "… goal-oriented thought and action in situations for which no routinised procedures are available" (Klieme et al., 2001:185, our translation).

This chapter describes recent results of problem-solving research with a focus on analytical problem-solving competence. From the perspective of a psychometric research approach, a key question is whether, apart from intelligence and other cognitive constructs, problem-solving competence represents a distinct construct, and if so, how it can be assessed and modelled. From an educational psychology perspective the focus lies on the question of how problem-solving competence can be fostered by instructional means.

In the context of school and education systems, research on problem-solving competence has become more relevant following the 2003 Programme for International Student Assessment (PISA) (OECD, 2003, 2005a). PISA 2003 defined cross-curricular problem-solving competence as "… an individual's capacity to use cognitive processes to confront and resolve real, cross-disciplinary situations where the solution path is not immediately obvious and where the literacy domains or curricular areas that might be applicable are not within a single domain of mathematics, science or reading" (OECD, 2003:156). The PISA problem-solving test showed some remarkable results, especially in Germany (Leutner et al., 2004; OECD, 2005a). While students in Germany only reached average results in mathematics, science and reading, their results in problem solving were above average (compared to the OECD mean of 500 score points; standard deviation = 100). This difference between students' problem-solving and subject-specific competencies, for example in mathematics, was especially pronounced in Germany, at 10 points on the PISA scale. Among all 29 participating countries only 2 countries showed larger differences in favour of problem solving: Hungary, at 11 points, and Japan, at 13 points (OECD, 2005a). This difference is particularly surprising against the background of the rather high correlation between problem-solving and mathematics scores (r = .89); the values for science were r = .80 and for reading r = .82 (OECD, 2005). According to the OECD report on problem solving in PISA 2003, this discrepancy can be interpreted in terms of a cognitive potential exploitation hypothesis: the cross-curricular problem-solving test reveals students' "… generic skills that may not be fully exploited by the mathematics curriculum" (OECD, 2005a:56; see also Leutner et al., 2004)

The cognitive potential exploitation hypothesis refers to the international part of the PISA 2003 test which contains analytical problems such as finding the best route on a subway map in terms of time travelled and costs (OECD, 2005a). These kinds of problems can be distinguished from dynamic problems such as a computer simulation of a butterfly swarm which has to be shaped in a certain way by using a controller whose functions are not mentioned in the problem description (Leutner et al., 2004; see also Greiff et al., 2012). In analytical problem solving, all the information needed to solve the problem is explicitly stated or can be inferred; it can thus be seen as the reasoned application of existing knowledge (OECD, 2005a). In dynamic problem solving, in contrast, most of the information needed to solve the problem has to be generated in an explorative interaction with the problem situation, or "learning by doing" (Wirth and Klieme, 2003). Dynamic problems are often characterised by a high number of interrelated variables (complexity), a lack of transparency regarding the causal structure and sometimes several, partly conflicting, goals and subgoals (Funke, 1991). In research, these types of problems are often referred to as complex problems (Funke and Frensch, 2007; Greiff et al., 2013) or, to use the terminology used in PISA 2012, as interactive problems (OECD, 2013, 2014; see also Chapter 6).

Results from a German PISA 2003 subsample show that the analytical and dynamic aspects of problem solving can be well distinguished empirically (Leutner et al., 2004). Analytical problem solving and subject-specific competencies in mathematics, science and reading form a comparably high correlated cluster, whereas dynamic problem solving seems to represent something different (see also Wirth, Leutner and Klieme, 2005). The disattenuated correlation between analytical and dynamic problem solving amounts to only $r = .68$ whereas analytical problem solving correlates with mathematics at $r = .90$ in this sample (Leutner et al., 2004). It can be shown that on the national PISA scale (Mean = 50; Standard deviation = 10) there is a lower mean difference between the lower vocational and the academic tracks of the German secondary school system for dynamic problem solving (14.7 points) than for analytical problem solving (19.0 points). This result points to a relative strength of students from the lower vocational track in dynamic problem solving, which can also be interpreted in terms of the cognitive potential exploitation hypothesis (Leutner et al., 2004). Despite these differences between analytical and dynamic problems, in PISA 2012, problem-solving competence was assessed using both analytical ("static" in PISA 2012 terminology) and interactive problems modelled on the same composite scale and labelled as creative problem solving (OECD, 2013, 2014; see also Chapter 5).

Finally, a subsample of the PISA 2003 students in Germany took part in a repeated measures study, which examined competence development in the 10th grade (from the end of 9th grade to the end of 10th grade) in mathematics and science (N = 6 020; Leutner, Fleischer and Wirth, 2006). A path analysis controlling for initial mathematical competence in 9th grade shows that future mathematical competence in 10th grade can be better predicted by analytical problem-solving competence ($R^2 = .49$) than by intelligence ($R^2 = .41$). In order to further explore the role of analytical problem solving for future mathematics, a communality analysis was conducted, decomposing the variance of mathematical competence (10th grade) into portions that are uniquely and commonly accounted for by problem-solving competence, intelligence and initial mathematical competence. Results showed that the variance portion commonly accounted for by analytical problem solving and initial mathematical competence ($R^2 = .127$) is considerably larger than the variance portion commonly accounted for by intelligence and initial mathematical competence ($R^2 = .042$) whereas the variance portion uniquely accounted for by both intelligence ($R^2 = .006$) and problem solving ($R^2 = .005$) are negligible (Leutner, Fleischer and Wirth, 2006). Further analysis including reading competence as an additional predictor does not change this pattern of results (Fleischer, Wirth and Leutner, 2009). These findings indicate that analytical problem-solving competence consists of several components with different contributions to the acquisition of mathematical competence. Although it is not entirely clear what these components are, it can be assumed that skills associated with a strategic and self-regulated approach to dealing with complex tasks with limited speed demands play an important role. Intelligence test items, in contrast, do not seem to require these skills, at least not to the same degree (Leutner et al., 2005).

In order to make use of the cognitive potential demonstrated by students in analytical problem solving for domain-specific instruction, three issues have to be dealt with. First, it is crucial to identify the cognitive requirements of analytical problems and the corresponding components of analytical problem-solving competence. Second, the interrelations of these components and their relations with external variables (such as subject-specific competencies) have to be investigated. This means establishing a structural model of analytical problem-solving competence and testing aspects of construct validity. Third, research is needed to determine how students' cognitive potential can be better exploited in domain-specific instruction. This requires developing and evaluating training methods in experimental settings which can then be transferred into school settings.

The cognitive potential exploitation hypothesis assumes that students have unused cognitive potential for science instruction as well as for mathematics (Rumann et al., 2010). In the following sections however, we will focus on the role of cross-curricular analytical problem solving for subject-specific problem solving in mathematics.

Analytical problem solving as a process

Based on early research on problem solving in the first half of the 20th century (e.g., Wertheimer, 1945), Polya (1945) describes problem solving as a series of four steps: 1) understanding the problem; 2) devising a plan; 3) carrying out the plan; and 4) looking back (evaluating the problem solution). In his model of problem solving, Mayer (1992) breaks the process of problem solving down into two major phases: 1) problem representation; and 2) problem solution. Problem representation can be further split into problem translation, in which the information given in the problem is translated into an internal mental representation, and problem integration, in which the information is integrated into a coherent mental structure. Likewise, the problem solution process can be divided into the two sub-processes of devising a plan of how to solve the problem and monitoring its execution.

These conceptualisations of the problem-solving process provide the basis for more recent approaches to both cross-curricular and subject-specific problem solving. The PISA 2003 assessment framework describes a comparable series of necessary steps in order to successfully solve the items of the (analytical) problem-solving test: 1) understand; 2) characterise; 3) represent the problem; 4) solve the problem; 5) reflect; and 6) communicate the problem solution (OECD, 2003:170 f.). A comparable series of steps can also be found in descriptions of the mathematical modelling cycle which constitutes the basis of the PISA mathematics test: 1) starting with a problem situated in reality; 2) organising it according to mathematical concepts; 3) transforming the real problem into a mathematical problem: generalising and formalising; 4) solving the mathematical problem; and 5) making sense of the mathematical solution in terms of the real situation (OECD, 2003:27; see also Freudenthal, 1983).

Analytical problem solving as a competence

Analytical problem solving and intelligence

In the context of research on cognitive abilities, problem solving is sometimes regarded as a component of intelligence (Sternberg and Kaufmann, 1998) or described as a cognitive activity associated with fluid cognitive abilities (Kyllonen and Lee, 2005). Problem solving can, however, be conceptually distinguished from intelligence: problem solving is characterised by a higher degree of domain specificity or situation specificity and it can be acquired through learning (Leutner et al., 2005). Furthermore, test items designed to tap the core domains of intelligence – general or fluid cognitive abilities – are generally decontextualised and less complex than analytical problem-solving test items, which always require a minimum level of content-specific knowledge (Leutner et al., 2005; see also Klieme, 2004). Results from PISA 2003, as well as results from studies on problem solving as an extension to PISA 2000 in Germany, provide further evidence that problem-solving competence and intelligence are structurally distinct. As shown by Wirth and Klieme (2003), students' scores on items developed to tap dynamic problem solving, analytical problem solving and intelligence are not adequately explained by a single latent factor, despite high correlations. Models with three latent dimensions provide a much better fit to the data (for a comparison between dynamic problem solving and intelligence see also Wüstenberg, Greiff and Funke, 2012).

Components of analytical problem solving

A number of cognitive components of problem solving can be identified from the problem-solving steps listed above, which are needed to successfully solve both cross-curricular analytical and subject-specific problems. Among these components are the availability and application of knowledge of concepts and knowledge of procedures (Schoenfeld, 1985; Süß, 1996). Knowledge of concepts is defined as knowledge of objects and states which can be used to construct an internal representation of the problem situation and the aspired goal state ("knowing that"; see Resnick, 1983). Knowledge of procedures is defined as knowledge of operations which can be applied to manipulate the problem situation, transforming it from its current state into the aspired goal state, as well as the ability to execute cognitive operations or procedures ("knowing how"; see Resnick, 1983). A further component is conditional knowledge which describes the conditions under which specific operations are to be applied ("knowing when and why"; Paris, Lipson and Wixson, 1983). Another important component is the availability and application of general problem-solving strategies which structure the search for relevant information, alternative problem representations, analogies or subgoals (Gick, 1986). Finally, self-regulation ensures that the processes of problem solving are properly planned, monitored, evaluated, and – if necessary – modified (Davidson, Deuser and Sternberg, 1994).

In a first attempt to examine the role of these components both for cross-curricular and subject-specific problem solving, Fleischer et al. (2010) analysed the cognitive requirement of the PISA 2003 problem-solving items and mathematics items. The aim of this analysis was to reveal the similarities and differences in the conceptualisation and construction of these test items with regard to the cognitive potential exploitation hypothesis (a similar analysis was conducted in order to compare problem solving and science items from PISA 2003; for details see Rumann et al., 2010). All PISA problem-solving items and a sample of the mathematics items (a total of 86 units of analysis) were evaluated by 3 expert raters using a classification scheme consisting of 36 variables (rating dimensions). Table 2.1 shows an excerpt of the results.

The results show, among other things, that the problem-solving items had a more detailed description of the given situation and a more personal and daily life context than the mathematics items. However, the problem-solving items do not clearly show the cross-curricular character which would have been expected from the PISA 2003 definition of problem solving. Even though most problem-solving items could not be associated with one particular subject area, approximately one-third of the items were rated as clearly mathematics-related. Mathematics items were characterised as more formalised and thus requiring greater ability to decode relevant information. This requires the availability and application of content-specific knowledge of concepts, which is also reflected in the higher level of the curricular knowledge needed to successfully solve the mathematics items. In case of problem-solving items, however, knowledge of procedures seems to be more relevant than knowledge of concepts. In particular, the problem-solving items placed greater requirements on students' ability to recognise restrictive conditions and to strategically plan the solution approach. This indicates that conditional knowledge and the ability to self-regulate are more relevant to finding the solution. These seem to be the relevant components of analytical problem solving rather than of mathematics competence (Fleischer et al., 2010).

Table 2.1 **Results of expert rating of PISA 2003 test items**

Rating dimension	Problem solving	Mathematics
Description of given situation	more detailed	less detailed
Item context	personal	global
Subject/domain area	30% mathematical	mathematical
Language	less formalised	more formalised
Recognising restrictive conditions	higher demands	lower demands
Strategic and systematic approach	higher demands	lower demands
Curricular-level of knowledge	low	high
Knowledge of concepts	lower demands	higher demands
Knowledge of procedures	higher demands	lower demands

Fleischer, J., J. Wirth and D. Leutner (2009). "Problem solving as both cross-curricular and subject-specific competence", paper presented at the 13th Biennal Conference of the European Association for Research on Learning and Instruction (Earli), Amsterdam, 29 August 2009.

Different dimensions of analytical problem solving

The PISA 2003 study assessed problem-solving competence using three different problem types: 1) decision making (e.g. finding the best route on a subway map in terms of time travelled and costs); 2) system analysis and design (e.g. assigning children and adults to dormitories in a summer camp considering certain rules and constraints); and 3) troubleshooting (e.g. finding out if particular gates in an irrigation system work properly depending on which way the water flows through the channels). However, the items were subsequently modelled on a one-dimensional scale.

Since these problem types differ with respect to the goals to be achieved, the cognitive processes involved and the possible sources of complexity (for a detailed description of the PISA 2003 problem-solving test items see OECD, 2005a), we looked more thoroughly into these dimensions of analytical problem-solving competence. We examined whether the different problem types used in PISA 2003 represent different dimensions of analytical problem solving, and whether these dimensions have differential relations with external variables (Fleischer, in prep.; see also Leutner et al., 2012). For that reason, the PISA 2003 problem-solving test was re-scaled based on the dataset from Germany (2 343 students with booklets with respective item clusters) according to the Partial Credit Model (Masters, 1982). We then compared the one-dimensional model with a three-dimensional model that allocates each item to one of three dimensions depending on the problem type (decision making, system analysis and design, or troubleshooting). The two models were evaluated by means of item-fit statistics (MNSQ-Infit), correlations of the dimensions of the three-dimensional model, reliabilities and variances of the dimensions, as well as the Bayes information criterion (BIC) and the likelihood ratio test (LRT) (see Wu et al., 2007). Finally, we estimated the correlations between the individual ability estimates from the three-dimensional model with external variables (school track, gender, subject-specific competencies and reasoning) in order to check for differential relations. Differential relations with these variables would provide a strong argument for a three-dimensional structure of analytical problem-solving competence.

While neither of the two estimated models could be preferred on basis of the item-fit statistics and the variances of the dimensions, the reliabilities showed a small advantage for the one-dimensional model (r = .78) compared with the three-dimensional model (r = .73; r = .73; r = .69). The dimensions of the three-dimensional model were highly correlated (r = .93; r = .83; r = .80), but not necessarily higher than the correlations that were usually found between competencies from different domains (OECD, 2005b). Results from the LRT showed a significantly better fit for the three-dimensional model than the more restricted one-dimensional model (ΔDeviance = $184_{(5)}$, $p < .001$), and the BIC also preferred the three-dimensional model (BIC_{3-dim} = 31 714 vs. BIC_{1-dim} = 31 859). Furthermore, the results displayed in Table 2.2 – although not entirely conclusive – provide some indications of differential relations of the dimensions of the three-dimensional model with external variables and therefore to some extent support the assumption of a three-dimensional structure of analytical problem-solving competence.

Table 2.2 **Intraclass correlation and relations with external variables for the three-dimensional model of analytical problem-solving competence: Mean differences and correlations**

	Decision making	System analysis and design	Troubleshooting
Intraclass correlation	.24	.32	.22
School track (reference = lower vocational track)			
Comprehensive track	0.341 [0.168/0.514]***	0.379 [0.156/0.602]**	0.405 [0.193/0.616]***
Higher vocational track	0.580 [0.446/0.714]***	0.645 [0.492/0.798]***	0.576 [0.438/0.714]***
Academic track	1.163 [1.046/1.280]***	1.293 [1.156/1.429]***	1.067 [0.949/1.184]***
R^2	.82	.76	.74
Gender (1 = female)	-0.045 [-0.117/0.027]	-0.024 [-0.090/0.042]	-0.149 [-0.230/-0.069]***
Mathematics	.611[.578/.644]***	.655[.624/.687]***	.568[.536/.600]***
Science	.582[.551/.612]***	.621[.584/.657]***	.530[.494/.567]***
Reading	.576[.544/.608]***	.608[.570/.646]***	.504[.469/.540]***
Reasoning	.464[.424/.503]***	.481[.439/.523]***	.394[.354/.434]***

Notes: Mean differences of person ability estimates of the three dimensional model for school track and gender. The intraclass correlation provides the proportion of overall between- and within-school variation that lies between schools. R2 provides the proportion of school level variance explained by school track. The person ability estimates are z-standardised weighted likelihood estimates. 95%-confidence intervals are in brackets. For mathematics, science, reading and reasoning the weighted likelihood estimates are correlated with each of the respective five plausible values from PISA 2003 (see OECD, 2005b), and then the mean of these five correlations is calculated per domain.

$^{***}p < .001$, $^{**}p < .01$. N = 2 343.

For all three dimensions, a substantial amount of the variance in the ability estimates can be allocated to the school level (between-group variance) but for the system analysis and design dimension this proportion (32%) is higher than for the other two dimensions. With respect to school track, all three dimensions show characteristic achievement differences, with the academic track having the highest mean difference (compared to the lower vocational track) followed by the higher vocational track and the comprehensive track. Even though the difference between the lower vocational track and the academic track is descriptively higher for the system analysis and design dimension (1.293) than for the other two dimensions (1.163 and 1.067), the overlapping confidence intervals prove this difference not to be significant. However, the school track variable explains more variance on the group level for the decision making dimension (82%) than for the other two dimensions (76% and 74%). Results by gender show a significant difference in favour of males (-0.149, CI95 = -0.230/-0.069) for the troubleshooting dimension. Furthermore, results for mathematics, science, reading and reasoning show substantial correlations with the person ability estimates for all three dimensions. Descriptively, these correlations are more pronounced for the system analysis and design dimension than the other two dimensions but examining the confidence intervals proves these differences to be significant only for the comparison between the system analysis and design and troubleshooting dimension (Fleischer, in prep.).

On the one hand, these three problem types seem to represent different but highly correlated dimensions, due to their rather different cognitive requirements. On the other hand, modelling analytical problem-solving competence as a one-dimensional construct in the context of large-scale assessments does not seem to cause a considerable loss of information. However, the multidimensional structure of analytical problem-solving competence outlined in the results above should be taken into account when assessing this competence on an individual level, especially when it comes to investigating the cognitive potential of analytical problem-solving competence in order to use it to foster subject-specific competencies, for example in mathematics.

Training in analytical problem-solving competence

The research results described so far indicate that analytical problem-solving competence, as it was assessed in PISA 2003, consists of different components which to some extent require different cognitive abilities. Therefore, systematic training in a selection of these components should have an effect on analytical problem-solving competence in general. In accordance with the cognitive potential exploitation hypothesis, and based on the assumption that both cross-curricular and subject-specific problem-solving competence share the same principal components, training in components of analytical problem-solving competence should also have transfer effects on mathematical competencies.

Based on the analysis of relevant components of analytical problem-solving competence and mathematics described above, three components – knowledge of procedures, conditional knowledge and planning ability (as part of self-regulation) – were chosen for two training studies (for details see Buchwald, Fleischer and Leutner, 2015; Buchwald et al., 2017).

Study I: Method

In a between-subjects design, a sample of 142 9th grade students (mean age 15 years; 44% female) was randomly assigned to one of two experimental conditions. The experimental group received 45minutes of computer-based multimedia training in problem solving focusing on knowledge of procedures, conditional knowledge and planning. The training was a mainly task-based learning environment with feedback. Students in the control group received a software tutorial dealing with an introduction to the graphical user interface of a computer-based geometry package.

Following the training phase, a post-test was administered with three parts: 1) test instruments assessing the three problem-solving components (knowledge of procedures, conditional knowledge and planning ability) as well as a global analytical problem-solving scale consisting of an item selection from the PISA 2003 problem-solving test; 2) test instruments assessing knowledge of procedures, conditional knowledge, and planning ability in the field of mathematics as well as a global mathematics scale consisting of an item selection from the PISA 2003 mathematics test; and 3) a scale assessing figural reasoning (Heller and Perleth, 2000) as an indicator of intelligence as a covariate.

We expected the experimental group to outperform the control group in all three components they had been trained in (treatment check) and on the global problem-solving scale. We further expected a positive transfer of the training on the three components in the field of mathematics as well as on the global mathematics test.

Study I: Results and conclusions

Due to participants' self-pacing in the training phase and the first part of the post-test, the control group spent less time on training and more time on the first part of the post-test than the experimental group – although pre-studies regarding time on task indicated equal durations of the experimental and control group treatment. Therefore, the first part of the post-test was analysed by means of an efficiency measure (post-test performance [score] per time [min]).

Analysis of (co-)variance controlling for school track and intelligence showed that the experimental group worked more efficiently on the items testing planning (η^2 = .073), conditional knowledge (η^2 = .200), and knowledge of procedures (η^2 = .020) than the control group, indicating a positive treatment check. The experimental group also outperformed the control group on the global problem-solving scale (η^2 = .026) However, regarding transfer on the mathematical scales, the results showed no significant differences between the performances of the two groups (η^2 = .005).

In conclusion, the problem-solving training proved to be successful at enhancing students' efficiency when working on items assessing knowledge of procedures, conditional knowledge and planning ability as components of analytical problem-solving competence. The training was also successful at raising students' efficiency when working on problem-solving items from PISA 2003, which further underlines the importance of these components for successful analytical problem solving in general. However, no transfer of training on the mathematical scales could be found (Buchwald et al., 2017).

Study II: Method

The second study implemented an extended problem-solving training with a longitudinal design, additional transfer cues and prompts to enhance transfer on mathematics. This field experimental study used 173 9th grade students (60% female; mean age 14 years) participating as part of their regular school lessons. The students in each class were randomly assigned to either the experimental or the control group and were trained in a weekly training session of 90 minutes. Including pre-test, holidays and post-test, the training period lasted 15 weeks. The experimental group received a broad training in problem solving with a focus on planning, knowledge of procedures, conditional knowledge and meta-cognitive heuristics. The control group received a placebo training (e.g. learning to use presentation software) consisting of exercises that were important within and outside the school setting.

The problem-solving and mathematics pre-test and post-test consisted of an item selection from the relevant PISA 2003 tests. All items were administered in a balanced incomplete test design with rotation of domains and item clusters to avoid memory effects between pre- and post-test. We expected the experimental group to outperform the control group on the problem-solving test (near transfer) as well as on the mathematics test (far transfer).

Study II: Results and conclusions

In order to investigate near transfer training effects, a linear model was calculated. The predictors were group membership and problem solving in the pre-test (T1), and the criterion was problem solving in the post-test (T2). This model showed a significant effect of problem solving at T1 (f^2 = .37) as well as a significant interaction of problem solving at T1 and group membership (f^2 = .22). To investigate far transfer training effects, a linear model was calculated with group membership and problem solving at T1, and mathematics at T1 as predictors and mathematics at T2 as criterion. This model showed significant effects for mathematics at T1 (f^2 = .55), problem solving at T1 (f^2 = .21) as well as a significant interaction effect of problem solving at T1 and group membership (f^2 = .23).

These aptitude treatment interaction (ATI) result patterns indicate the extended problem-solving training to be effective for both near and far transfer for students with low initial problem-solving competence, but not for those with high initial problem-solving competence (Buchwald, Fleischer and Leutner, 2015; Buchwald et al., 2017).

Summary and discussion

The development of problem-solving competence is a crucial goal of various educational systems (OECD, 2013) from two perspectives: first, problem-solving competence is an important outcome of educational processes and second, it is also an important prerequisite for successful future learning both in and out of school. The German results from the PISA 2003 study of cross-curricular (analytical) problem solving suggested that students in Germany possess generic skills or cognitive potential which might not be fully exploited in subject-specific instruction in mathematics (OECD, 2005a; Leutner et al., 2004),

To investigate this hypothesis and to make use of students' cognitive potential for subject-specific instruction, it is necessary to gain a better understanding of the cognitive requirements of the PISA 2003 problem-solving items and their relationship to other cognitive constructs. This understanding could be used to develop and evaluate training and instructional tools aimed at improving problem-solving competence.

There are strong similarities between the steps and cognitive resources needed to solve mathematics and problem-solving items. Knowledge of concepts, knowledge of procedures, conditional knowledge, general problem-solving strategies, and the ability to self-regulate are all to different degrees relevant for both problem solving and mathematics (Fleischer et al., 2010). The three problem types used in PISA 2003 – decision making, system analysis and design, and troubleshooting – represent distinct but highly correlated dimensions of problem-solving skills with somewhat different relations to external variables such as subject-specific competencies and reasoning ability. The substantial correlations of these dimensions with subject-specific competencies, particularly for mathematics, suggest that a training in some aspects of problem solving should also improve subject-specific skills. Results from the German PISA 2003 repeated measures study also support this assumption (Leutner, Fleischer and Wirth, 2006).

A first experimental study into training on the components of knowledge of procedures, conditional knowledge and planning ability (as part of self-regulation) showed improved efficiency on the trained components and improved problem solving in general. This result further underlines the relevance of these components for analytical problem-solving competence, although the training failed to show transfer effects in mathematics (Buchwald et al., 2017). A second training study using an extended problem-solving training programme showed some evidence for the cognitive potential exploitation hypothesis. Among low-achieving problem solvers the training improved both problem-solving and mathematical competence, but not among students with high initial problem-solving competence, indicating a compensatory training effect (Buchwald, Fleischer and Leutner, 2015; Buchwald et al., 2017). These results are in line with the interpretation of the German PISA 2003 results that emphasised unused cognitive potential especially for low-achieving students (Leutner et al., 2005).

Students in Germany performed above the OECD average in both problem solving and mathematics in PISA 2012 (OECD, 2014). However, since the test focus has changed from analytical problem solving in PISA 2003 to creative problem solving combining both analytical and complex problems in PISA 2012 (OECD, 2013, 2014), future research has to elaborate whether this result indicates a better exploitation of students' cognitive potential. Future research should also address potential reciprocal causal effects between problem solving and mathematics in order to determine how cross-curricular problem-solving competence could be used to foster mathematical competence (Leutner et al., 2004; OECD, 2005a) and vice versa (Doorman et al., 2007; OECD, 2014).

References

AAAS (American Association for the Advancement of Science) (1993), *Benchmarks for Science Literacy*, Oxford University Press, New York.

Blum, W. et al. (eds.) (2006), *Bildungsstandards Mathematik: Konkret (Mathematics Educational Standards: Concrete)*, Cornelsen, Berlin.

Bodner, G.M. and J.D. Herron (2002), "Problem-solving in chemistry", in J.K. Gilbert et al. (eds.), *Chemical Education: Towards Research-Based Practice*, Kluwer, Dordrecht, pp. 235-266.

Buchwald, F., J. Fleischer and D. Leutner (2015), "A field experimental study of analytical problem solving competence – Investigating effects of training and transfer", *Thinking Skills and Creativity*, Vol. 18, pp. 18-31. http://dx.doi.org/10.1016/j.tsc.2015.04.009.

Buchwald, F. et al. (2017), "Training of components of problem-solving competence – An experimental study on the cognitive potential exploitation hypothesis", in D. Leutner, J. Fleischer, J. Grünkorn and E. Klieme (eds.), *Competence assessment in education: Research, models and instruments, pp. 315-331*, Springer, Berlin.

Charles, R.I. and E.A. Silver (eds.) (1988), *The Teaching and Assessing of Mathematical Problem Solving*, National Council of Teachers of Mathematics, Reston, VA.

Davidson, J.E., R. Deuser and R.J. Sternberg (1994), "The role of metacognition in problem solving", in J. Metcalfe and A.P. Shimamura (eds.), *Metacognition: Knowing about Knowing*, MIT Press, Cambridge, MA, pp. 207-226.

Dörner, D. (1976), *Problemlösen als Informationsverarbeitung (Problem Solving as Information Processing)*, Kohlhammer, Stuttgart.

Doorman, M. et al. (2007), "Problem solving as a challenge for mathematics education in the Netherlands", *ZDM Mathematics Education*, Vol. 39/5, pp. 405-418. http://dx.doi.org/10.1007/s11858-007-0043-2.

Duncker, K. (1935), *The Psychology of Productive Thinking*, Springer, Berlin.

Fleischer, J. (in preparation), "Strukturen fächerübergreifender analytischer Problemlösekompetenz" (Structures of cross-curricular analytical problem-solving competence), unpublished doctoral dissertation, University of Duisburg-Essen, Essen, Germany.

Fleischer, J., J. Wirth and D. Leutner (2009). "Problem solving as both cross-curricular and subject-specific competence", paper presented at the 13th Biennal Conference of the European Association for Research on Learning and Instruction (Earli), Amsterdam, 29 August 2009.

Fleischer, J. et al. (2010), "Strukturen fächerübergreifender und fachlicher Problemlösekompetenz – Analyse von Aufgabenprofilen" (Structures of cross-curricular and subject-related problem solving competence – Analysis of task profiles), *Zeitschrift für Pädagogik*, Vol. 56, pp. 239-248.

Freudenthal, H. (1983), *Didactical Phenomenology of Mathematical Structures*, Reidel, Dordrecht.

Funke, J. (1991), "Solving complex problems: Human identification and control of complex systems", in R.J. Sternberg and P.A. Frensch (eds.), *Complex Problem Solving: Principles and Mechanisms*, Erlbaum, Hillsdale, NJ, pp. 185-222.

Funke, J. and P.A. Frensch (2007), "Complex problem solving: The European perspective – 10 years after", in D.H. Jonassen (ed.), *Learning to Solve Complex Scientific Problems*, Erlbaum, New York, pp. 25-47.

Gick, M.L. (1986), "Problem-solving strategies", *Educational Psychologist*, Vol. 21/1-2, pp. 99-120.

Greiff, S., S. Wüstenberg and J. Funke (2012), "Dynamic problem solving: A new assessment perspective", *Applied Psychological Measurement*, Vol. 36/3, pp. 189-213, http://dx.doi.org/10.1177/0146621612439620.

Greiff, S., S. Wüstenberg, G. Molnár, A. Fischer, J. Funke and B. Csapó (2013), "Complex problem solving in educational settings – Something beyond *g*: Concept, assessment, measurement invariance, and construct validity, *Journal of Educational Psychology*, Vol. 105/2, pp. 364-379, http://dx.doi.org/10.1037/a0031856.

Heller, K.A. and C. Perleth (2000), *Kognitiver Fähigkeitstest (Cognitive Ability Test)*, Beltz, Göttingen.

Klieme, E. (2004), "Assessment of cross-curricular problem-solving competencies", in J.H. Moskowitz and M. Stephens (eds.), *Comparing Learning Outcomes: International Assessments and Education Policy*, Routledge, London, pp. 81-107.

Klieme, E. et al. (2001), "Problemlösen als fächerübergreifende Kompetenz? Konzeption und erste Resultate aus einer Schulleistungsstudie" (Problem solving as cross-curricular competence? Concepts and first results from an educational assessment), *Zeitschrift für Pädagogik*, Vol. 47/2, Beltz, Weinheim, pp. 179-200.

Kyllonen, P.C. and S. Lee (2005), "Assessing problem solving in context", in O. Wilhelm and R.W. Engle (eds.), *Handbook of Understanding and Measuring Intelligence*, Sage, Thousand Oaks, CA, pp. 11-25.

Leutner, D., J. Fleischer and J. Wirth (2006), "Problemlösekompetenz als Prädiktor für zukünftige Kompetenz in Mathematik und in den Naturwissenschaften" (Problem solving competence as a predictor of future competencies in mathematics and science), in M. Prenzel et al. (eds.), *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres*, Waxmann, Münster, pp. 119-137.

Leutner, D. et al. (2012), "Analytisches und dynamisches Problemlösen im Lichte internationaler Schulleistungsvergleichsstudien. Untersuchungen zur Dimensionalität" (Analytical and dynamic problem solving from an international educational studies perspective: Analysis of dimensionality), *Psychologische Rundschau*, Vol. 63/1, pp. 34-42. http://dx.doi.org/10.1026/0033-3042/a000108.

Leutner, D. et al. (2005), "Problemlösefähigkeit als fächerübergreifende Kompetenz" (Problem solving as cross-curricular competence), in E. Klieme, D. Leutner and J. Wirth (eds.), *Problemlösekompetenz von Schülerinnen und Schülern*, VS, Wiesbaden, pp. 11-19.

Leutner, D. et al. (2004), "Problemlösen" (Problem solving), in PISA-Konsortium Deutschland, *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs*, Waxmann, Münster, pp. 147-175.

Masters, G.N. (1982), "A rasch model for partial credit scoring", *Psychometrika*, Vol. 47/2, pp. 149-174. http://dx.doi.org/10.1007/BF02296272.

Mayer, R.E. (1992), *Thinking, Problem Solving, Cognition*, Freeman, New York, New York.

Mayer, R.E. and M.C. Wittrock (2006), "Problem solving", in P.A. Alexander and P.H. Winne (eds.), *Handbook of Educational Psychology*, Erlbaum, Mahwah, NJ, pp. 287-303.

NCTM (2000), *Principles and Standards for School Mathematics*, National Council of Teachers of Mathematics, Reston, VA.

OECD (2014), *PISA 2012 Results: Creative Problem Solving (Volume V): Students' Skills in Tackling Real-Life Problems*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264208070-en.

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

OECD (2005a), *Problem Solving for Tomorrow's World: First Measures of Cross-Curricular Competencies from PISA 2003*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264006430-en.

OECD (2005b), *PISA 2003 Technical Report*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264010543-en.

OECD (2003), *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264101739-en.

Paris, S.G., M.Y. Lipson and K.K. Wixson (1983), "Becoming a strategic reader", *Contemporary Educational Psychology*, Vol. 8/3, pp. 293-316. http://dx.doi.org/10.1016/0361-476X(83)90018-8.

Polya, G. (1945), *How to Solve It*, University Press, Princeton, New Jersey.

Resnick, L.B. (1983), "Toward a cognitive theory of instruction", in S. Paris, G. Olson and H. Stevenson (eds.), *Learning and Motivation in the Classroom*, Erlbaum, Hillsdale, NJ, pp. 5-38.

Rumann, S. et al. (2010), "Vergleich von Profilen der Naturwissenschafts- und Problemlöse-Aufgaben der PISA 2003-Studie" (Comparison of profiles of science and problem solving tasks in the PISA 2003-study), *Zeitschrift für Didaktik der Naturwissenschaften*, Vol. 16, pp. 315-327.

Schoenfeld, A.H. (1985), *Mathematical Problem Solving*, Academic Press, Orlando, FL.

Sternberg, R.J. and J.C. Kaufman (1998), "Human abilities", *Annual Review of Psychology*, Vol. 49, pp. 479-502.

Süß, H.-M. (1996), *Intelligenz, Wissen und Problemlösen (Intelligence, Knowledge, and Problem Solving)*, Hogrefe, Göttingen.

Törner, G., A.H. Schoenfeld and K.M. Reiss (2007), "Problem solving around the world: Summing up the state of the art", *ZDM – The International Journal on Mathematics Education*, Vol. 39/5-6, pp. 353-353. http://dx.doi.org/10.1007/s11858-007-0053-0.

Wertheimer, M. (1945), *Productive Thinking*, Harper & Row, New York.

Wirth, J. and E. Klieme (2003), "Computer-based assessment of problem solving competence", *Assessment in Education: Principles, Policy and Practice*, Vol. 10/3, pp. 329-345 http://dx.doi.org/10.1080/0969594032000148172.

Wirth, J., D. Leutner and E. Klieme (2005), "Problemlösekompetenz – Ökonomisch und zugleich differenziert erfassbar?" (Problem solving competence – Measurable parsimoniously but comprehensively?), in E. Klieme, D. Leutner and J. Wirth (eds.), *Problemlösekompetenz von Schülerinnen und Schülern. Diagnostische Ansätze, theoretische Grundlagen und empirische Befunde der deutschen PISA–2000-Studie*, VS, Wiesbaden, pp. 73-82.

Wu, M.L. et al. (2007), *ACER ConQuest Version 2.0: Generalised Item Response Modelling Software*, ACER Press, Camberwell, Victoria.

Wüstenberg, S., S. Greiff and J. Funke (2012), "Complex problem solving – More than reasoning?", *Intelligence*, Vol. 40/1, pp. 1-14, http://dx.doi.org/10.1016/j.intell.2011.11.003.

Chapter 3

Problem solving: Understanding complexity as uncertainty

By
Magda Osman
Experimental Cognitive Psychology, Department of Psychology, Queen Mary University of London, United Kingdom.

Understanding complexity has dominated the domain of problem solving. There needs to be a cohesive idea about what makes problem solving hard in contexts broadly described as complex, and what are the sources of errors in problem solving in situations thought to be complex. This chapter argues that our understanding of complexity requires an overhaul, and perhaps a more encouraging and fruitful way forward in research on complex problem solving, is to reconceptualise "finding solutions to complex problems" as instead "findings ways of reducing uncertainty". This entails redefining the domain of "complex dynamic problem solving" into "decision making under uncertainty". This enables a Bayesian approach to understanding what makes a situation controllable, or appear to be controllable to a problem solver, and hence what cognitive and psychological factors they bring to bear on the situation.

Introduction

The Programme for International Student Assessment (PISA) project published a report in 2005, *Problem Solving for Tomorrow's World*, which identified problem solving as a key component for future assessment. Why? The report's argument, which applies equally to its studies in the domain of cognitive science, is that the value and importance of assessing problem-solving ability, is because it is a practical skill (OECD, 2005).

Typically the cognitive operation of problem solving refers to any type of goal-directed set of actions or cognitive activities (Anderson, 1982) designed to move the problem solver from a state in which the problem is identified and represented, to an end state in which manipulations of components of the problem are carried out in order to find a solution. Of course, it is also important to accurately represent the situation to facilitate the process of finding a solution, as well as to find the most appropriate strategy to move from the current to the end state (Jonasson, 2000).

The ability to identify and represent a problem accurately, as well to plan the sequence of steps to reach a solution is necessary in virtually all our personal and professional lives (Simon et al., 1987; Jonasson, 2000; Osman, 2010a). To investigate how we typically go about solving problems, and finding ways of training people to improve their skills, the cognitive science domain has developed at least six classes of problem-solving task: insight, information-search, analogical, mathematical, scientific thinking and complex dynamic. Each generally captures abstract aspects of everyday situations that involve combining actions in the most efficient (and often creative) way possible to reach an end state (for example getting three missionaries across a river without them being eaten by cannibals). One of these six branches that has garnered much attention in the PISA project and has been incorporated into the broad assessment criteria is complex dynamic problem solving.

The first aim of this chapter is to set out precisely what complex dynamic problem solving is, with illustrations from some everyday situations, as well as empirically notable examples. Through this it should become apparent: 1) why complex dynamic problem solving has had an established research tradition; 2) why the PISA project has identified it as an important component in its assessment criteria; and 3) why there is growing demand for researchers to communicate their findings more widely (Levy and Murnane, 2006). The second aim of this chapter is a little more challenging. Complex problem solving has a long tradition in psychology, and what is more, its reach extends to other disciplines, such as computer science, engineering, human factors, management, economics and other social sciences. The term "complex" is in and of itself complex. It has also been used to refer to problems that typically invoke a form of thinking in the problem solver that requires a complex sequence of actions. The main problem has been that, while researchers in the field of problem solving tend to notionally understand what it is that makes a problem complex, there is no agreed definition of what it is for a task to be complex, and no formal basis on which to distinguish complex from non-complex problems.

Thus, the second part of this chapter presents the case for approaching this type of problem from the viewpoint that the tasks developed to study complex dynamic problem solving actually invoke a form of decision making under uncertainty. To establish the argument, this section comments briefly on the differences between decisions made under risk, uncertainty and ambiguity. It presents a Bayesian framework that can be used to conceptualise the way people judge a situation, and uses a decision theory approach to establish the basis on which people then choose to act.

One might ask: why bother replacing the term "complex" with "uncertain"? The chapter responds in more detail later but, in short, there is a limit to how much we can understand and improve on the skills needed to tackle everyday situations if we don't have an agreed way of defining the situations in the first place. In general, the view taken here is that uncertainty can come from having to juggle multiple aspects of situation that need to be considered all at once – for example offsetting the side effects of drug treatment with the benefits of stemming the spread of a disease (i.e. a vast hypothesis space) – or from simply not knowing all of the aspects that need to be considered in the situation that one faces – for example buying an apartment in an economic downturn. There can be many more sources of uncertainty than just these two, but they all reduce to a subjective judgment that an individual makes as to the controllability of a situation, from which they then decide what action to take.

In this light, the assessments of complex problem solving in the student cohorts studied in the PISA project, and more generally in the cognitive science domain, might best be thought of instead as assessing people's ability to successfully and reliably reduce uncertainty.

Complex problem solving: Everyday examples

As part of a morning ritual, the doctor has advised us to take an aspirin every day. So, dutifully we take our aspirin from the medicine cabinet every morning, and replace it in its rightful spot routinely every day. On one occasion the box of aspirin is missing from the usual spot in the cupboard. We check the locations we visited in the house after we took the aspirin, and find that we'd put the box in our work bag, as a reminder to replace it because it is running empty. Aside from being mundane, and rather dull, this example represents a problem (locating the box) with a single goal (taking the aspirin regularly), and a simple solution: search the last known locations until the box is found (there are of course other simple alternative solutions). What also makes this example seem trivially simple is that it is a stationary problem: we couldn't find the box because we hadn't put it in its usual place; it is extremely unlikely that there are external factors that will alter the location of the box other than our own actions.

As soon as we enter into the realm of non-stationary problems things become increasingly difficult, and far less trivial. Climate change is one of the most critical issues that we currently face. Consequently, there have been various government initiatives in many countries encouraging people to moderate their fuel consumption. So, now imagine that we have decided to take action and become more energy conscious. As a result, we have installed an energy monitoring device (smart meter) to help reduce us track our energy consumption. This example represents a problem (work out the best course of action that will help economise home fuel consumption). There are now two competing goals (reliably reducing energy consumption while at the same time maintaining a standard of living that doesn't compromise all members of the household). The solution is also non-trivial (for example, working out the energy consumption of the appliances in the home and adapting our behaviour accordingly to reach an optimal balance). Our initial strategy is to start small by asking all members of the household to switch all appliances off at the mains after usage, recharge phones in the evening only and to shower five minutes less each day. Our smart meter shows that since installing it, for the period of time that we have been deliberately economising, we have successfully lowered our energy consumption, and moreover our fuel bill has marginally reduced. But, since this is a non-stationary problem, the strategy has to be flexible, because we know that things will change from one day to the next, such as individual demands, price structures and seasonal changes.

For instance, we would need to take into account the number of appliances used at any one time, the actually time the appliances are used and the frequency with which they are used. In addition, there will be other factors that change, such as the daily and seasonal differences in the demand for using appliances from different members of the household. More to the point, we need to evaluate our success relative to a target, otherwise we won't know how effective our strategies are over time, and relative to others. Therefore, we need to face a further issue of what target to choose. If we settle that issue, we face a further concern which is how to interpret the information we are getting back from the smart meter, which will be changing over time and only providing a gross indication of energy usage but, crucially, not how much energy each specific appliance uses over a specific amount of time. It is clear from just this case, which is a general everyday problem that households face, just how detailed and difficult a non-stationary problem can be.

Complex problem solving: Empirical examples

While not the earliest example (Toda, 1962), or the most empirically valid version (for example Dörner, 1975; Funke, 1992), one of the most often cited, and most famous examples of a complex dynamic problem-solving task was developed by Berry and Broadbent (1984, 1987, 1988). In their task, participants were presented with a story in which they are told that they are managers in a sugar factory. Their main goal is to figure out precisely how many workers should be employed at any one time to keep sugar production at a specified level. Participants need to track: 1) the amount of sugar actually produced (i.e. achieved outcome); 2) the amount of sugar that needs to be produced (i.e. target outcome); and 3) the number of workers they have chosen to employ at any one time (i.e. action). What makes this task non-static is reflected in the underlying rule that connects up these three variables. The task operates based on the following rule: $(P = (2 * W – P_{1t\text{-}1}) + R)$ in which the relevant variables are the workforce (W), current sugar output (P), previous sugar output (P1) and a random variable (R). The decisions of the participants in one trial will influence the outcome of the next. This means that the outcome will actually change, although in this case, it is still anchored to a choice that the participant must make. The addition of a random variable also obscures the relationship between decision and outcome, which makes the relations uncertain and of course harder to learn.

Compared with the everyday smart meter example, even though there are so few elements that need to be attended to or recalled, and even though only one element needs to be manipulated at any one time, Berry and Broadbent's task is still hard for essentially two reasons. In part, the difficulty comes from the ambiguity associated with uncovering the precise relationships between one's actions and their effects. The other reason is that participants are faced with two tasks at once: 1) learning about the relationship between actions and their effects; and 2) exploiting that relationship in order to reach and maintain a specific goal. Both these reasons are essentially the same critical issues that make everyday problem solving hard (Osman, 2010a).

A more recent complex dynamic problem-solving task which, like Berry and Broadbent's, also includes a random variable as a method of making it non-static, was developed by Osman and Speekenbrink (2011, 2012). In this task, participants are told that they are part of a medical research team conducting drug trials, and that their goal was to learn to reach and maintain a specified level of neurotransmitter ("N") release. They could achieve this by manipulating the amount of hormone from a choice of three (Hormone "A", "B" and "C") injected into a patient suffering from severe stress. By manipulating different levels of different hormones at different times they could learn the relationship between their actions and the effects on the outcome. More formally, the task environment can be described as in the following equation: $y(t) = y(t\text{-}1) + .65\ _{x1}(t) - .65\ _{x2}(t) + e(t)$, in which $y(t)$ is the outcome on trial t, $x1$ is the positive variable (Hormone A), $x2$ is the negative variable (Hormone B) and e a random noise component, normally distributed with a zero mean and standard deviation of 4 (low noise) or 16 (high noise). The $x3$ null variable (Hormone C) is not included in the equation as it had no effect on the outcome. Here the variables that participants need to track are: 1) the amount of neurotransmitter actually released (i.e. achieved outcome); 2) the amount of neurotransmitter that needs to be released (i.e. target outcome); 3) the number of hormones injected at any one time (i.e. action); and 4) the level of each hormone injected at any one time (i.e. action).

As with Berry and Broadbent's task, the outcome changes as a result of the participant's actions. However, even if the participant chooses to do nothing on a trial, the outcome will change anyway as a result of the way in which the random variable is included in the rule. Another difference between the two tasks is that in Osman and Speekenbrink's (2011) task, participants need to attend to, recall and directly manipulate more components at any one time. One might argue that because of these factors this task more closely approximates the everyday smart meter example discussed previously. But, how can one know?

It seems intuitive that one task (Berry and Broadbent, 1984) is less complex than the other (Osman and Speekenbrink, 2011). But what does that mean? In what way is one more complex than the other? Is it in terms of the task itself and its structure, or is it psychologically more complex? Clearly these issues resonate with any assessment exercise which needs to compare people's ability in a variety of increasingly difficult problem-solving tasks. One needs to know how they increase in difficultly, and if this occurs on the same identifiable dimensions. These issues also matter hugely in applied contexts. Imagine now that your smart meter not only gives you up-to-the-minute readings of your home energy usage, but also forecasts the likely usage of energy for the rest of the day with a detailed breakdown by appliance that could be used. By adding more layers of information onto existing devices such as smart meters or the systems people interact with from day to day, what new problem-solving situations will people face? Without knowing how to describe the contexts in which people have to perform complex dynamic problem solving, we won't know what the effects on problem solving will be when the contexts change.

Complexity by any other name: Uncertainty

As highlighted in the previous section, there is a need to characterise problems that seem to intuitively vary according to some form of complexity. In fact, uncovering the objective characteristics of a complex problem that make it complex has been a preoccupation in the study of complex problem solving, but to no avail (Campbell, 1988). Knowing the features that could be used to define a problem as complex has not helped to advance our understanding of why it is that people generally tend to falter in some problems and not others, or for that matter, why it is that some people are able to learn to control situations reliably better than other situations (Quesada et al., 2005). Psychological studies suggest that objective characteristics of complexity (the number of variables in the problem and the type of relationships between variables – i.e. non-linearities, and delays between a manipulated variable and an outcome) do not reliably predict our ability to surmount complex dynamic problems. They also cannot be used to accurately gauge the accuracy of people's representation and understanding of the problem (Osman, 2010a, 2010b).

In recent years, this has prompted a shift in research focus onto problem-solving behaviour (e.g. strategies, learning styles, control behaviours and transferability of skills) and examining the effects of psychological factors (e.g. self-doubt, anxiety, fear, misperception/misrepresentation of the problem) on problem solving (Bandura, 2001; Campbell, 1988; Gonzales et al., 2005; Sterman, 1994). This has been an alternative route to understanding complexity, by defining it in relation to the psychological factors that can lead to deterioration in problem-solving behaviour. I will argue here that in order to advance our understanding of the differences in the contexts in which dynamic problem solving is examined, and the effects produced on problem-solving behaviour, problem-solving researchers need to reconceive complexity in terms of uncertainty. In previous work in complex dynamic problem-solving tasks, I and others have shown that people form judgments of uncertainty while problem solving and the judgments correspond with, and can accurately track, changes in rates of learning and performance while solving a problem (Brown and Steyvers, 2005; Osman, 2008; Vancover et al., 2008; Yeo and Neal, 2006; Yu, 2007). One of the cues that people use when judging the uncertainty of a problem is the rate at which a non-stationary problem is likely to change over time (Osman, 2010a). They integrate this estimate with estimates of how confident they are that they can predict the changes occurring, and how confident they are that they can control the changes occurring. The following discussion focuses on demonstrating the role of uncertainty in the context of complex dynamic problem solving.

The connection between uncertainty and problem solving is by no means new. Take the following comment by Jonasson:

> "Just what is a problem? There are only two critical attributes of a problem. First, a problem is an unknown entity in some situation (the difference between a goal state and a current

> state). Those situations vary from algorithmic math problems to vexing and complex social problems, such as violence in the schools. Second, finding or solving for the unknown must have some social, cultural, or intellectual value. That is, someone believes that it is worth finding the unknown. If no one perceives an unknown or a need to determine an unknown, there is no perceived problem (whether the problem exists independent of any perception is an ontological issue that is beyond the scope of this paper). Finding the unknown is the process of problem solving." (Jonasson, 2000:65)

As Jonasson claims, at the heart of what makes a problem an actual problem is the concept of the "unknown", and what needs to be surmounted in order to solve the problem is the process by which people transform the unknown to the known (i.e. problem solving). While the tasks that have been developed to examine complex dynamic problem solving are not uniform (in other words they cannot be made equivalent in terms of structure, or content), they reduce to one thing, which is that they present people with a context in which they are uncertain about what generates an outcome (i.e. the unknown), and they are tested on reaching a goal which reflects their ability to generate a particular outcome on successive occasions (surmounting the unknown).

In the study of decision making, a distinction is often made between decision making under risk, and decision making under uncertainty. I would argue that complex dynamic problem solving is an example of decision making under uncertainty. Although Knight's (1921) classic distinction was originally presented within the context of economics, the distinction between risk and uncertainty which concerns an agent's source of knowledge regarding outcomes and probabilities extends well beyond economics into psychology and neuroscience, in which it is often used. Knight distinguished between 1) *a priori* probabilities, which can be logically deduced, as in games of chance; 2) statistical probabilities, derived from data; and 3) estimates, arising from situations in which "there is no valid basis of any kind for classifying instances" (Knight, 1921/2006:225). Decision making under risk refers to situations in which people are engaged in planning actions against knowledge of the probabilities of the outcomes following their actions (i.e. *a priori* probabilities come to bear in these contexts). Decision making under uncertainty concerns situations in which people plan their actions from limited available knowledge of the possible outcomes, and in which the probabilities of the outcomes following actions is not known, or cannot be known, i.e. statistical probabilities and estimates come to bear in these contexts (Osman, 2011).

If we return to the examples discussed in the previous section, the objective that the problem solver is trying to achieve throughout is to find ways in which the cues can be manipulated by whichever means necessary in order to reach and maintain a certain outcome; in non-stationary environments the aim is to achieve this reliably over time. However, in all but one of the examples (i.e. the aspirin example) problem solvers have to find a way of acting in a situation with either incomplete or unreliable information. This can be compounded because the problem solver can be faced with ill-defined problems (Simon, 1973), for which there are no optimal or well-defined solutions, or in which the potential solutions to the problem can generate conflicts between choices of actions that serve competing goals. Regardless of the source of uncertainty (the number of variables that need to be manipulated, the underlying relationship between variables etc…), the uncertain will pervade because in complex dynamic problems the outcomes are non-deterministic, and the problem itself is non-stationary (for example the sugar factory task or the medical decision-making problem). Therefore, the problem solver cannot ever be sure how much the changes that are generated are based on their own actions and how much are based on the inherent workings of the problem situation itself (i.e. the underling endogenous properties of the system), or a combination of both (Osman, 2010a, 2010b).

Clearly then, from this characterisation, the everyday and empirical examples of complex dynamic problem solving fit within the definition of decision making under uncertainty (Knight, 1921; Osman, 2011). While researchers into complex problem solving might not be able to provide a comprehensive and systematic way of identifying complexity along a single dimension, or for that matter mapping complexity onto problem-solving performance, I would argue that it is possible to treat all complex dynamic control tasks as equivalent on the basis of their subjective degree of controllability, by reconceptualising complexity in terms of uncertainty. Thus, complex dynamic control tasks now lend themselves to a Bayesian approach in which researchers can examine what makes a situation more or less controllable, given the various sources of uncertainty (e.g. problem solving context, problem solver) that contribute to making an outcome dynamic.

The Bayesian approach proposed by Harris and Osman (2012) suggests that problem solvers evaluate situations from which they make assessments of controllability in two steps: first, they make a judgment about the controllability (or lack of controllability) of a situation, and second, they make a decision to act according to this judgment. The judgments are informed by the available evidence from the situation combined with a prior degree of belief in the controllability of the situation. This decision theory framework suggests that decisions are made following probabilistic judgments about the controllability of a situation. Harris and Osman (2012) proposed that the decision will be influenced by the utility (subjective goodness or badness) associated with misclassifications of controllable events as uncontrollable (misses/the illusion of chaos) and uncontrollable events as controllable (false positives/the illusion of control; see Table 3.1). Asymmetries in people's utilities enhance the likelihood for an objectively uncontrollable situation to be rationally, although wrongly (from an objective perspective), classified as controllable (the illusion of control), and explain why people act in objectively controllable situation as if they are uncontrollable (the illusion of chaos).

Table 3.1 **The four possible outcomes of decisions based upon the controllability of a situation**

		State of the world	
		Controllable	Uncontrollable
Act as though the world is…	Controllable	Hit	False positive (illusion of control)
	Uncontrollable	Miss (illusion of chaos)	Correct rejection

Source: Harris and Osman (2012), "Illusion of control: A reflection of biased perceptions of random outcomes or prior expectations of good and bad outcomes".

Cues to controllability

Both prescriptive frameworks set out precisely how to judge the controllability of a situation and how to act from that. To begin, the Bayesian framework represents degrees of belief as single event probabilities (or distributions over these probabilities), and people have a degree of belief in the truth of the hypothesis that "Outcome X is controllable". Bayes' theorem sets out the way in which people should update their degree of belief in a hypothesis (e.g. that an outcome is under their control) from evidence they receive:

$$P(\mathrm{h}\backslash\mathrm{e}) = \frac{P(\mathrm{h})\ P(\mathrm{h}\backslash\mathrm{e})}{P(\mathrm{e})}$$

Here, a problem solver's posterior degree of belief in the hypothesis, given a piece of evidence (from the problem solving situation itself), $P(h|e)$, is a function of their degree of belief in the hypothesis before receiving evidence (the prior), $P(h)$, the likelihood of the evidence given the truth of the hypothesis, $P(e|h)$, and the base rate of the evidence, regardless of the truth status of the hypothesis, $P(e)$. Prior beliefs will be informed by the extent to which they have found themselves in situations that are unfamiliar from which they were then able to control the outcome. For instance, to return to the home energy monitoring device (the smart meter). The father in the family may never have encountered a smart meter before, but he may have been used to using other self-monitoring devices for many years (blood sugar monitor, blood pressure monitor, pedometer). His experience with other monitoring devices may lead to his judgement of a high likelihood of controllability. Moreover, the problem-solving context will provide evidence. That is, the situation itself will also contain cues as to the controllability of the device, for example the simplicity of the interface, the number of buttons on the device, the comprehensibility of the instructions manual and the amount of information presented on screen.

Now imagine presenting the same individual with a smart fitness monitoring device which tracks intake (calories) and calories spent minute by minute through daily activities (e.g. walking, running, sleeping). How much more or less complex is this problem-solving situation as compared to managing one's energy consumption? While it might be hard to align both problems in the context of what might be deemed "objective task characteristics", by looking at the subjective judgments people make as to the controllability of both the smart meter and the fitness meter examples, we can place the two on the same dimension, which makes them comparable to each other. The advantage of empirical contexts is that we can manipulate complex dynamic problem-solving tasks in such a way as to understand more precisely the kind of evidence (i.e. cues to control) that people use to inform their judgments in combination with their prior experience of controllable situations. In two recent reviews Osman (2010a, 2010b) examined the findings from a literature that has been steadily amassing since the 1960s across diverse disciplines (economics, management, psychology, engineering, computer science, human factors and neuroscience). The list of criteria below is condensed from over 300 research articles on complex dynamic problem solving. Moreover, these aren't only the cues that people use to identify the controllability of a situation; these are also the criteria that people use to ensure that a highly uncertain non-stationary situation is made controllable.

1) **A consistent goal, and faith in the control system:** The person must have a clear and stable objective and hold a reasonable belief that their actions (through the control system) can affect the thing being controlled.

2) **A system fit for purpose:** The control system itself has a single, stable, overarching function, which coheres with the aims of the individual using it.

3) **A predictable system:** The control system must be (to some extent) predictable. For this, it must change at a manageable rate (not too rapidly or "violently").

4) **Time and space for control:** A series of sequential steps (that follow a linear order) are required, and these must happen in a particular time and place.

5) **A coherent system:** There is sequencing and synchronisation of the core operations in the system (i.e. the various components will work effectively together to further the core function of the control system). No part of a control system can function in a way that conflicts with the main objective of the control system.

6) **Feedback and controlled adjustment:** The information in the system feeds back on itself, and allows for adaptive adjustment to feedback (but with a threshold – so that adjustments do not threaten the overall objective of the control system).

The second component of the assessment of the problem-solving context is whether or not to act, which is simply a binary choice. Converting a ratio scale (i.e. a probabilistic belief in the controllability of a situation) to a simple binary choice needs a threshold value to be set. Harris and Osman (2012)'s application of decision theory to the context of control sets out that the output reflects both degrees of belief (probabilities) and the utility (subjective goodness or badness) of different possible outcomes. In any situation, there are two possible states of the world: the situation is controllable or the situation is not controllable. Similarly, one can act as though the world is controllable or uncontrollable. Thus, there are four potential outcomes as presented in Table 3.1.

In decision theoretic terms, there are two classifications, those which correctly correspond to the world as controllable or uncontrollable (hits and correct rejections) and those which are incorrect with respect to the world (false positives and misses). Positive utility is associated with correct classifications and negative utility is associated with incorrect classifications. Harris and Osman (2012) argue that if the positive utility associated with a hit equals the positive utility associated with a correct rejection, and the negative utilities associated with false positives and misses match, then the threshold for deciding that a situation is controllable or uncontrollable is: P(*controllable*) = .5. As soon as there is an asymmetry in the utilities, however, then the threshold moves from .5. Thus, if one perceives higher costs with a miss than with a false positive, the threshold will be less than .5, increasing the likelihood of acting as though a situation is controllable.

Importantly, judgments of controllability will change, as will decisions to act, because the problem the solver is trying to resolve is a non-stationary one. Therefore, judgments of the controllability of the situation change as the problem solver continues to interact with the situation, and so will their decisions to act as if is the situation is controllable or not. Therefore, ongoing evaluations of performance in complex dynamic problem solving contexts should not only involve accuracy and reliability in control, but also ongoing judgments of control (Osman, 2008) and decisions of control.

Practical solutions to practical problems

At the heart of this article has been a pragmatic issue, and that is the need to understand the skills needed for effective problem solving, and the criteria that we use to define the problem, both in science and in the assessment exercise in PISA designed to enhance learning. In the domain of "complex dynamic problem solving", which I argue should be referred to as decision making under conditions of uncertainty, we are still very far off from tackling these issues. However, it is very bad form to end on a negative note. So, the parting message from this article is that there are ways of reconceptualising complex problems into uncertain situations. The advantage in doing so from an empirical perspective is that we can feasibly make a variety of different types of problems equivalent. This is an important first step because they can then be directly compared on the same dimension, and the dimension combines aspects of the objective characteristics of the problem and the psychological factors that problem solvers bring to bear. The simple solution to a very long and entrenched problem is to focus on the subjective judgments that people about the controllability of the problem-solving context. This may help to better understand an essential cognitive activity that has also been identified as key in a major worldwide assessment exercise.

References

Anderson, J.R. (1982), "Acquisition of cognitive skill", *Psychological Review*, Vol. 89, pp. 369-406.

Bandura, A. (2001), "Social cognitive theory: An agentic perspective", *Annual Review of Psychology*, Vol. 52, pp. 1-26.

Berry, D. C. and D.E. Broadbent (1988), "Interactive tasks and the implicit-explicit distinction", *British Journal of Psychology*, Vol. 79/2, pp. 251-272.

Berry, D.C. and D.E. Broadbent (1987), "The combination of explicit and implicit learning processes in task control", *Psychological Research*, Vol. 49/1, pp. 7-15.

Berry, D.C. and D.E. Broadbent (1984), "On the relationship between task performance and associated verbalizable knowledge", *Quarterly Journal of Experimental Psychology*, Vol 36/2, pp. 209-231.

Brown, S. and M. Steyvers (2005), "The dynamics of experimentally induced criterion shifts", *Journal of Experimental Psychology: Learning, Memory & Cognition*, Vol. 31/4, pp. 587-599.

Campbell, D. (1988), "Task complexity: A review and analysis", *Academy of Management Review*, Vol. 13/1, pp. 40-52.

Dörner, D. (1975), "Wie Menschen eine Welt verbessern wollten und sie dabei zerstörten" [How people wanted to improve a world and in the process destroyed it], *Bild der Wissenschaft*, Vol 12/2, pp. 48-53.

Funke, J. (1992), "Dealing with dynamic systems: Research strategy, diagnostic approach and experimental results", *German Journal of Psychology*, Vol. 16/1, pp. 24-43.

Gonzales, C., P. Vanyukov and M.K. Martin (2005), "The use of microworlds to study dynamic decision making", *Computers in Human Behavior*, Vol. 21/2, pp. 273-286.

Harris, A. and M. Osman (2012), "The illusion of control: A Bayesian perspective", *Synthese,* Vol. 189/1, pp. 29-38.

Jonasson, D. (2000), "Toward a design theory of problem solving", *Educational Technology Research and Development*, Vol. 48/4, pp. 63-85.

Knight, F.H. (1921/2006), *Risk, Uncertainty and Profit*, Houghton Mifflin, Boston.

Levy, F. and R. Murname (2006), "How computerized work and globalization shape human skill demands", *IPC Working Paper Series*, MIT-IPC-05-006, Industrial Performance Center, Massachusetts Institute of Technology, Cambridge, MA.

OECD (2005), *Problem Solving for Tomorrow's World: First Measures of Cross-curricular Competencies from PISA 2003*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264006430-en.

Osman, M. (2011), "The role of feedback in decision making", in A.Q. Rana (ed.), *Diagnosis and Treatment of Parkinson's Disease*, InTech Publishers.

Osman, M. (2010a), *Controlling Uncertainty: Decision Making and Learning in Complex Worlds*, Wiley-Blackwell Publishers

Osman, M. (2010b), "Controlling uncertainty: A review of human behavior in complex dynamic environments", *Psychological Bulletin*, Vol. 136/1, pp. 65-86.

Osman, M. (2008), "Positive transfer and negative transfer/antilearning of problem-solving skills", *Journal of Experimental Psychology: General*, Vol.137/1, pp. 97-115.

Osman, M. and M. Speekenbrink (2012), "Prediction and control in a dynamic environment", *Frontiers in Psychology*, Vol3/68, pp. 1-11. http://dx.doi.org/10.3389/fpsyg.2012.00068.

Osman, M. and M. Speekenbrink (2011), "Cue utilization and strategy application in stable and unstable dynamic environments", *Cognitive Systems Research*, Vol. 12/3-4 pp. 355-364.

Quesada, J., W. Kintsch and E. Gomez (2005), Complex problem-solving: a field in search of a definition?, *Theoretical issues in ergonomics science*,Vol. 6(1), pp.5-33.

Simon, H.A. (1973), "The structure of ill structured problems", *Artificial Intelligence*, Vol. 4/3-4, pp. 181-201.

Simon, H.A. et al. (1987), "Decision making and problem solving", *Interfaces*, Vol. 17/5, pp. 11-31.

Sterman, J.D. (1994), "Learning in and about complex systems", *System Dynamics Review*, Vol. 10/2-3, pp. 291-330.

Toda, M. (1962), "The design of the fungus-eater: A model of human behavior in an unsophisticated environment", *Systems Research and Behavioral Science*, Vol. 7/2, pp. 164-183.

Vancouver, J.B., K.M. More and R.J. Yoder (2008), "Self-efficacy and resource allocation: Support for a nonmonotonic, discontinuous model", *Journal of Applied Psychology*, Vol. 93/1, pp. 35-47.

Yeo, G.B. and A. Neal (2006), "An examination of the dynamic relationship between self-efficacy and performance across levels of analysis and levels of specificity", *Journal of Applied Psychology*, Vol. 91/5, pp. 1088-1101.

Yu, A. (2007), "Adaptive behavior: Humans act as Bayesian learners", *Current Biology*, Vol. 17/22, pp. R977-R980.

Chapter 4

Problem solving from a mathematical standpoint

by
John A. Dossey
Distinguished Professor of Mathematics Emeritus, Illinois State University, United States.

Problem solving has always been at the heart of both pure and applied mathematics. This chapter examines the evolution of the concept of what constitutes a problem, and the models and findings concerning problem solving from research into the learning and teaching of mathematics. It then considers the emerging understanding of the role played by metacognition in current conceptions of problem solving in mathematics and its learning.

Introduction

Mathematics has been described as a search for regularity in patterns, dimension, quantity, uncertainty, shape and change. As a mathematician abstracts what might be a pattern, he or she forms a conjecture of what might constitute a description of that pattern. This then shifts the focus to testing whether related generalisations may be verified. Such searching is at the heart of problem solving. In the abstracting of information, one is engaged in understanding, selecting strategies and representations, and viewing both mentally and figuratively a multitude of instances and potential cases to form a statement describing the problem and its context. In the verification, one is striving to find a solution (Steen, 1990).

While problem solving is at the heart of mathematics, there is little agreement in the field about what problem solving is, what its boundaries are, or how it is best learned or taught to students in schools or practised in adult life. Gardiner indicates that the word "problem" has a negative connotation suggesting an "unwanted and unresolved tension." On the other hand, "solving" has the notion of "a release of tension" (2008: 995). The juxtaposition of these terms throughout mathematics and mathematics education literature suggests that 1) there must be some known way of releasing the tension; and 2) that this process is something that works and can be taught to those wishing to solve problems. While mathematicians and mathematics educators have suggested various methods and approaches, their success in transferring these practices across varied domains within mathematics is questionable. Gian-Carlo Rota, the famed combinatorialist, wrote: "Why don't we tell the truth? No one has the faintest idea how the process of scientific induction works and by calling it a 'process' we may be already making a dangerous assumption" (1986:263). This caution aside, this chapter examines some of the models and practices proposed by mathematicians, mathematics educators, psychologists and other students of problem solving.

Mathematicians' views of problem solving

While mathematicians from the time of Plato and Aristotle have made comments on the development of mathematical ideas and the values of deductive and inductive approaches, we begin with the comments of the French mathematician Jacques Hadamard (1945) describing his introspective study of problem solving, as well as the results from a survey he made of his contemporaries.

Hadamard begins by refuting the notion that there is a "mathematical brain" and that those endowed with one have special powers to create in mathematical venues. He notes "we shall see that there is not just one single category of mathematical minds, but several kinds, the differences being important enough to make it doubtful that all such minds correspond to one and the same characteristic of the brain" (1945: 5). Turning from this, Hadamard examines the preoccupation of psychologists with notions of sudden chance discoveries or discoveries as the result of lengthy periods of systematic, logical reasoning. Mathematicians, in a survey conducted by M. Edmond Maillet, a colleague of Hadamard, discount chance discoveries of major relationships, as well as ideas springing forth from a dream (Allen, 1888). However, they did ascribe discoveries to having worked hard on a problem, putting it aside for a while, only to return again and then be able to make new progress. Among the responses, there was support for the role that emotions play in problem solving, a topic we will return to later.

Hadamard spent a good deal of time analysing Henri Poincaré's comments on his discoveries based on a presentation he made at the Société de Psychologie in Paris in 1937 (Poincaré, 1952). Poincaré noted several instances of working diligently on problems, setting them aside, and embarking on other tasks, only to have the solution appear to him unexpectedly as he was engaged in other activities. Similar reports are attributed to Gauss, Helmholtz and other members of the scientific community. Hadamard contrasts this sudden clarification of prior efforts from the unconsciousness with the claims of sudden initial discoveries of intact generalisations.

As a result of his study, Hadamard proposed a four-stage model for mathematical problem solving. It consists of a period of preparation, a period of incubation, a period of illumination and a period of verification. The first period, that of preparation, consists of the framing of the question through reading, initial work, analysis of partial results and failed attempts, and conscious examination of potential strategies which may further the work. The second period, that of incubation, consists of the acts of the unconscious mind in combining possibilities and even, potentially, ranking them in terms of "beauty and usefulness" (1945: 32). The third period, that of illumination, is perhaps the least understood or developed. However, it is the stage through which the varied possibilities are brought forward to conscious thought in a more organised and interpretable way than had every before been considered or formulated. Once these new possibilities are on the table, the conscious mind can then assist in the "précising" of these new avenues of attack and the fourth period, that of verification, during the establishment of related solutions to the core problem that initiated the work in the first place. Many mathematicians ascribe to similar events in their work toward finding the solution(s) to problems, both working alone and when working with and on the discoveries of others.

George Pólya, the famous Hungarian mathematician and author of *How to Solve It* (1945), took the discussion of problem solving another step forward in proposing a similar four-step method accompanied by a large number of examples. Pólya's contribution was aimed at teachers and students and shifted the consideration of problem solving from strictly pure mathematical or mathematical physics problems to a wider class of mathematics contexts. In it, Pólya popularised the notion of heuristics or specialised methods or strategies that work across classes of problems. Pólya's four-step method was framed by the stages of understanding the problem, devising a plan, carrying out the plan and looking back.

In *How to Solve It* the discussion of mathematical problem solving is presented alongside a series of examples and problem-solving strategies, or heuristics, related to mathematics education settings. Pólya states that "Routine problems, even many routine problems, may be necessary in teaching mathematics but to make the students do no other kind is inexcusable. Teaching the mechanical performance of routine mathematical operations and nothing else is well under the level of the cookbook because kitchen recipes do leave something to the imagination and judgment of the cook but mathematical recipes do not" (1945: 172). Pólya's model is perhaps the best known of the models proposed by mathematicians. Further, its focus, through examples, on a variety of heuristics, both named these approaches and set them up as targets of instruction in teaching "problem solving" to students as a core process in school mathematics.

What is a problem?

To this point, we have proceeded as if everyone is in agreement about what a problem actually is. Some mathematicians, as in Gian-Carlo Rota's earlier quote, note the extreme difficulty that exists in communicating the differences between real problem solving and working exercises using a specific strategy. This is the difficulty that surfaces in reducing problem solving to textual material and in attempting to assist teachers to come to recognise, internalise and model problem solving in the mathematical sense in their classrooms. This difficulty in delineating what a "problem" is clouds communication with teachers and the public about the difference between practising routine exercises and engaging in real "problem solving".

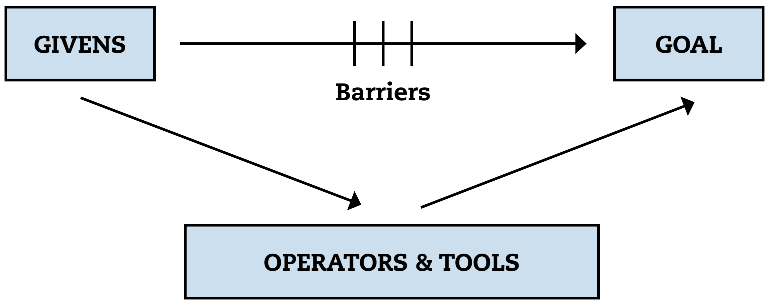
What is unique about a mathematical problem is that it involves a situation, posed in either an abstract or contextual setting, where the individual wrestling with the situation does not immediately know how to proceed or of the existence of an algorithm that will immediately move toward a solution. Further, there is a lingering sense of "unknowingness" about the situation. However, the statement or a representation of the problem situation is within the individual's ken.

Problem solving requires those being asked to participate in searching for a solution to be capable of understanding what the problem itself is. From a mathematical standpoint, this brings up one necessary condition. The problem solver must either know the relevant concepts' definitions or be capable of finding and deciphering them should they not know them. Further, the problem solver needs to understand how to structure mathematical generalisations and representations in such a way as to communicate his or her understanding and the path to a solution or a set of solutions. Further, the individual has to be willing and capable of actually participating in the search for a solution.

This said, let us return to the consideration of what a "problem" actually is. A "problem" exists when an individual has a particular goal, but doesn't know how to achieve it (Duncker, 1945). Problems can range from the task of finding what number to add to 29 to get a sum of 73 for a five-year-old, to resolving the Riemann conjecture for a professional mathematician. In both cases, there is a barrier between the problem solver's current knowledge and the desired result. The presence of this barrier helps conceptually structure the nature and understanding of a problem situation. In either of these cases, there is a lack of a critical connection based on conceptual knowledge or the lack of an appropriate algorithmic approach that stymies the problem solver. In other cases, it may be the lack of knowledge of a theorem (a principle or generalisation), an emotional block or lack of access to tools that creates the barrier between the solver and the solution. As a result of one or more of these impediments, the situation may remain and the problem stays unsolved. Depending on a student's motivation – another source of barriers – the individual may or may not return to try to solve the problem at another time (Funke, 2010).

Figure 4.1 provides a visual representation of the nature of a problem, with the givens and goal state shown with barriers preventing an immediate solution. The alternative path, avoiding the barriers, provides a route to the goal via operators (selections leading to relevant information from heuristic strategies, new representations, expert assistance) and information from tools, such as computer software, calculation devices and measurement instruments (Mayer, 1990; Frensch and Funke, 1995; Mayer and Wittrock, 2006; OECD, 2005).

Figure 4.1 **Problem Situation**



Source: Frensch, P.A. and J. Funke (eds.) (1995), Complex Problem Solving: The European Perspective, Erlbaum Associates, Hillsdale, NJ.

The OECD's Programme for International Student Assessment (PISA) structured its definition of problem-solving competency for its 2012 assessment of 15-year-olds' problem-solving capabilities as follows:

> Problem-solving competency is an individual's capacity to engage in cognitive processing to understand and resolve problem situations where a method of solution is not immediately obvious. It includes the willingness to engage with such situations in order to achieve one's potential as a constructive and reflective citizen. (OECD, 2013: 122)

The last sentence in the OECD's conception of problem-solving competency is recognition that competencies are important factors in how individuals interact with and shape situations they encounter in their lives. It reflects that a person competent in solving problems approaches a problem in a reflective and self-monitored, responsible way. This suggests the role that metacognition plays in problem solving competence, a topic that will be discussed later.

The role of problem solving in the development of mathematics

Problems have played a role in mathematics from the beginning of recorded history. In fact, some of the oldest mathematical artefacts are semblances of modern day mathematical textbooks, with examples of algorithms and comments on their application to the solution of problem-like situations. One of the oldest is the Ahmes, or Rhind, papyrus located at the British Museum. It consists of some 87 mathematical problems written for developing problem-solving competency in students of the period circa 1650 BC (Gillings, 1972). The mathematics education of individuals over successive centuries was guided in large by similar "texts" that provided algorithmic procedures and then practised their application to solving "problems" representative of the use of mathematics in the learners' culture.

However, at the turn of the 20th century, mathematicians like Felix Klein in Germany, John Perry in England and Eliakim Hastings Moore in the United States issued a call for a shift in emphasis, providing a greater focus on developing individual thinking skills and lessening the emphasis on rote memory and practice in the learning of mathematics. Others issued concerns for the development of these process skills to be balanced with some continued practice of more traditional procedural skills (Stanic and Kilpatrick, 1989).

Stanic and Kilpatrick divide the focus on problem solving surfacing in the learning of mathematics into three roles: problem solving as context, problem solving as skill and problem solving as art. The first of these roles sees problem solving as providing contexts that serve as means to other ends. These means include justifying the study of mathematics itself, motivating learners, providing recreation to encourage students, introducing a mathematical concept or skill, or providing practice in a newly encountered algorithm or concept. The second role sees problem solving in the learning of mathematics as an end goal of the act itself. However, this role can sometimes have a dark side in practice, as it leads to a segregation of problems into routine and non-routine problems, with the emphasis in learning often shifting to an increased focus on routine problems. The third role for problem solving is that of problem solving as an art. This interpretation embraces the work of Pólya and other proponents of problem solving who differentiate between active reasoning and plausible reasoning, as Pólya liked to term it (Pólya, 1954). He differentiated active reasoning from the finished reasoning presented in polished, formal deductive proofs. He embraced the use of active reasoning through the inspection of students immersed in a good problem, working their way through towards a solution, working individually or in groups. Further, he urged teachers to assume the roles of posers of problems, of supportive guides providing motivation, of questioners that caused students to reflect on their work, and as co-investigators with their students.

Key to this third conception of problem solving and the related active roles for teachers are two distinctions. The first is between exercises and problems. To Pólya, and most mathematicians and mathematics educators, the key feature of a problem is that one does not have an immediate plan of what to do. This conception of problem solving shows it as an interactive process cycling between the known and the unknown.

The second distinction is between a list of problem-solving strategies and problem-solving itself. Schoenfeld (1992) provides a thoughtful analysis of the multitude of often-contradictory conceptions of problem-solving instructional practices and their implementation in contemporary classrooms. In summarising the research, Begle had the following summary of the state of the field:

> A substantial amount of effort has gone into attempts to find out what strategies students use in attempting to solve mathematical problems…. No clear-cut directions for mathematics education are provided by the findings of these studies. In fact, there are enough indications that problem solving strategies are both problem- and student-specific often enough to suggest that finding one (or few) strategy which should be taught to all (or most) students are far too simplistic. (1979:145)

While pessimistic, Begle's summary was directly on point. Many statements about problem-solving strategies had been reduced to algorithmic recipes for teaching problem solving, often illustrated with examples that were, at best, exercises for the level of students for which the recipes were being touted. Writing a decade later, Dossey and colleagues found little had changed:

> Instruction in mathematics classes is characterized by teachers explaining material, working problems at the board, and having students work mathematics problems on their own - a characterization that has not changed across the eight-year period from 1978 to 1986.
>
> Considering the prevalence of research suggesting that there may be better ways for students to learn mathematics than listening to their teachers and then practicing what they have heard in rote fashion, the rarity of innovative approaches is a matter for true concern. Students need to learn to apply their newly acquired mathematics skills by involvement in investigative situations, and their responses indicate very few activities to engage in such activities. (Dossey et al., 1988:76)

Subsequent retrospective examinations of students' and teachers' reports of learning activities and overall student performance have failed to change this view of classroom instruction and the time devoted to achieving real problem-solving capabilities in students in United States classrooms. Information from other countries suggests that although substantial time is allocated to problem solving, students' capabilities have not yet reached the levels of expectations held by their school programmes (Törner, Schoenfeld and Reiss, 2007).

Students' problem solving as viewed through PISA

The PISA 2003 special assessment of cross-curricular problem solving (OECD, 2005) presented 15-year-old students in 40 countries with 10 problem situations that together contained a total of 19 tasks. These tasks were spread across three types of problem-solving contexts:

- **Decision making:** Decision-making items required students to choose among alternatives under a system of constraints. The major task was to understand the goal state and the constraints acting on the alternatives, then make a decision and evaluate its effectiveness.
- **System analysis and design:** System analysis and design items required students to identify relationships between parts of a system and/or design a system to express the relationships between parts. The major task was to identify the relevant parts of the system and the relationships among them and their relationship to the goal and then to analyse or design a system that captures the relationships and to check or evaluate that it does.
- **Troubleshooting:** Troubleshooting items required students to diagnose and correct a faulty or underperforming system or mechanism. The major task was to understand the features of the system and their relationship to the malfunction, including causal relationships. The solver needs to diagnose the malfunction, propose a remedy and evaluate the implementation of the remedy.

The assessment was presented in a paper-and-pencil assessment. The tasks ranged from multiple-choice items to items requiring students to construct and communicate their solution to the problem situation presented. The complete set of problems and comparative data are available for downloading via OECD (2005).

The scale resulting from students' performances described four distinct levels of problem solvers. The PISA problem-solving scale was set so that the mean student performance across OECD countries, weighted equally, resided at 500 points with a standard deviation of 100 score points. Students scoring below 405 points were described as Below Level 1: weak or emergent problem solvers. Students in this class failed to understand even the easiest items in the assessment or failed to apply the necessary processes to characterise important features or representations of problems. At best, they could only deal with straightforward tasks or make observations requiring few or no inferences.

Students scoring from 405 points up to 499 points are described as Level 1: basic problem solvers. These students could solve problems based on only one data source containing discrete, well-defined information. They understand the nature of problems and can find and organise information related to the major features of a problem. Then they can transfer from one representation to another if that transformation is not too complex, as well as check a number of well-defined conditions within the problem.

Students scoring from 500 points up to 591 points were classed as Level 2: reasoning, decision-making problem solvers. These students could use analytic reasoning skills and solve problems requiring decision making. They could apply inductive and deductive reasoning, reason about cause and effects, and reason with several information combinations including cases requiring a systematic comparison of all possible cases. They were able to combine and synthesise information from a variety of sources and representations in moving to solutions.

Students scoring 592 points and up were labelled as Level 3: reflective, communicative problem solvers. Students at this level not only analyse problems and make decisions, but they are also capable of thinking about the underlying relationships and how they relate to the solution of the problem. Further, they approach their work in a systematic fashion, constructing their own representations along the way to solving and verifying their solutions. These students are capable of communicating the solution to others in writing and through the use of varied representations. Level 3 problem solvers are capable of co-ordinating a large number of conditions, monitoring variables and temporal conditions, and moving back and forth with changes in interconnected variables. Perhaps the most striking aspect about the work of these students is their organisation and ability to monitor their work with a focus on developing a clear and coherent solution to the problem at hand (OECD, 2005).

While a full accounting of student performance can be found in OECD (2013), suffice it to say that the story of the distribution of problem-solving capabilities is in the variation in performance. About half of the students in the OECD PISA 2003 assessment countries and partner economies score at or above Level 2. Seventy percent or more of the students in Finland, Japan, Korea and Hong Kong, China score at Level 2 or above. Less than 5% of the students in Indonesia and Tunisia were able to meet the same criterion. Further, more than one-third of the students in Japan and Hong Kong, China perform at Level 3. Korea had the highest mean score (550) for individual countries and partner economies, though it was not significantly higher than the scores of the trio of Hong Kong, China (548); Finland (548) and Japan (547).

Large-scale assessments such as PISA provide some gross comparisons of countries' student performance and reflect, to some degree, the emphasis placed on aspects of problem solving in their learning experiences. Most enticing from these results, however, is the fact that the higher one goes in the performance rankings of students in large-scale assessments or smaller studies found in master's or doctoral theses, the more self-organisation and monitoring begin to stand out as criteria differentiating students' performance in problem-solving settings. It is this capability to self-monitor and regulate one's resources in problem solving we turn to next. There were similar findings from the results of the PISA 2012 problem-solving assessment which involved computer-delivered complex problems requiring a greater degree of active student engagement with the problems than the 2003 PISA problem-solving assessment. Further chapters in this work will focus on the results of the 2012 PISA problem-solving study.

The role of metacognition in mathematical problem solving

A common theme running through studies of students' problem-solving behaviour is the observation that successful problem solvers have control over their progress and rely on reflective pauses to organise their observations, abstract potential patterns, form summary statements and then plan additional data gathering or attempts to develop a communication about their solution (Silver, Branca and Adams, 1980; Schoenfeld, 1985). These plans often parallel the behaviour of experts in acquiring the knowledge they need. This observed behaviour is referred to as metacognition. Bransford, Brown and Cocking (1999) define this behaviour as "the ability to monitor one's current level of understanding and decide when it is not adequate … Adaptive experts are able to approach new situations flexibly and to learn throughout their lifetimes. They not only use what they have learned, they are metacognitive and continually question their current levels of expertise and attempt to move beyond them" (pp. 47-48). A great deal of literature has been developed to indicate the role such metacognitive reflections and monitoring play in overcoming students' initial misconceptions and in developing strong conceptual models that gradually replace those misconceptions with solid habits of mind that form a basis for successful problem solving and mathematical inquiry. However, one should not assume that all successful problem solvers would approach a problem in the same fashion, or that the process of understanding and analysing a problem can be made algorithmic (Schoenfeld, 1985, 1989, 1992).

While the exact link between problem solving and mathematical thinking is unknown, various models have been proposed to account for their mutual support of one another. In general the metacognitive models tend to focus on the following five factors (Schoenfeld, 1992; Lester, Garofalo and Kroll, 1989):

- the knowledge base
- problem-solving strategies
- monitoring and control
- beliefs and affects
- instructional practices.

Based on these factors, proposed by Schoenfeld, or similar models, Schoenfeld and others in mathematics education have used them to examine mathematical thinking and problem solving.

The knowledge base

First and foremost in examining problem solving is the amount and nature of knowledge the solver brings to the situation in which the problem solving is to be focused or the interactions a person has with other individuals or environments. Some researchers also focus on the joint roles of actions and their specific embedding in the environment in which the problem is embedded, arguing for the role of situated cognition (Brown et al., 1989; Lave and Wenger, 1991). A third aspect that Schoenfeld discusses is the role that representations (verbal, figural, graphical, tabular and symbolic) play in establishing one's knowledge base (Schoenfeld, 1992).

Schoenfeld argues that the knowledge base factor breaks down into two areas. "What information relevant to the mathematical situation or problem at hand does he or she possess? And how is that information accessed and used?" (1992: 349). The first area speaks to the facts, concepts, principles, strategies, algorithms, beliefs and attitudes the solver brings to the table. Central in this is the logical differences between facts, definitions, concepts, principles, algorithms and strategies and the differing roles they play in exploring contexts, relating information, ordering hypotheses and verifying assertions on the way to solving a problem. The navigation of this information requires the development of a personal epistemology of mathematical knowledge and the development of categories within it. The second area speaks to the structure of memory (Silver, 1987).

One theory is that the long-term memory is structured as a neural network whose vertices represent major chunks of memory and those edges represent connections between those chunks. Here again epistemology comes into play in considering the types of knowledge. Gilbert Ryle (1949) divided knowledge into “knowing that” and “knowing how”. Cooney and Henderson (1971), looking at mathematics content, divided knowledge into deductive, inductive, classifying and analysing, as a means of assisting students to relate different forms of knowledge and apply it in different situations. In another analysis, Henderson (1967) examined how instruction might link aspects of concepts to aid in understanding or structuring instruction. Others have suggested other instances and models for dividing knowledge into scripts, frames or schemata. All indicate that problem solvers tend to find some organisational form for problem contexts and that these forms help govern their behaviour when they encounter similar experiences in the future.

Problem-solving strategies

As already discussed, there are a number of models that have been developed by mathematicians and mathematics educators to codify approaches to problems and their solutions. Pólya (1945) suggested broad strategies of analogy, auxiliary elements, decomposing and recombining, induction, specialisation, variation, and working backward. However, these have been hard to implement on a general scale, even though they have held up with more mathematically inclined audiences.

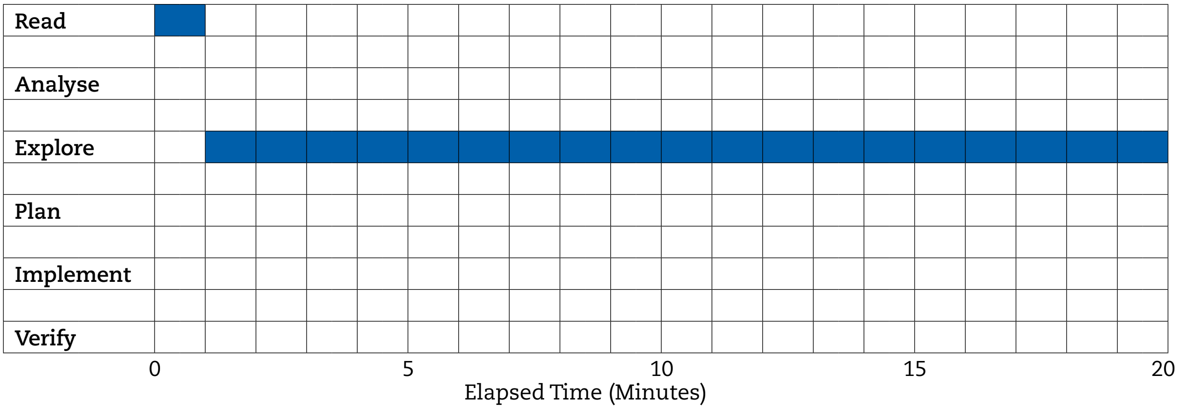
Studies have indicated that, given strategic guidance in structuring learning sequences and focused work with the strategies, combined with enough time, teachers can separate themselves from the “sage on the stage” role to that of “consultant and guide.” (Charles and Lester, 1984). This study and others reflect how difficult it is to implement the power of teachers who can model problem-solving strategies as Pólya envisioned, even with excellent problems and outstanding teachers.

Self-regulation, or monitoring and control

Self-regulation, or monitoring and control, is at the heart of metacognition and, as such, at the heart of mathematical problem solving. Self-regulation is that sixth sense that expert problem solvers have about stopping to check progress or recognising that one needs to change their conceptual framework, representation or tool usage in considering a particular piece of mathematics. The mind is indicating that their present course is not quite as smooth as it should be, or that there is a vital missing link in the explanation being developed. This monitoring is essential in the overall evaluation progress and the selection of the next step in problem solving. Silver, Branca and Adams (1980), Lesh (1985), and Garfalo and Lester (1985) were among the first mathematics educators to see the important role that metacognition played in problem solving, in the form of self-regulation and monitoring.

Perhaps the best illustration of the role of such self-regulation is in the time-effort graphs Schoenfeld (1989) developed from observations of a pair of students, a mathematician and a pair of students’ solving problems after finishing a course in problem solving. The following graphs reflect the activities of problem solvers during intervals of time devoted to attempts to solve problems. Figure 4.2 shows the work of a pair of students attempting to solve a problem.

Figure 4.2 **A pair of students' problem-solving activities over time**

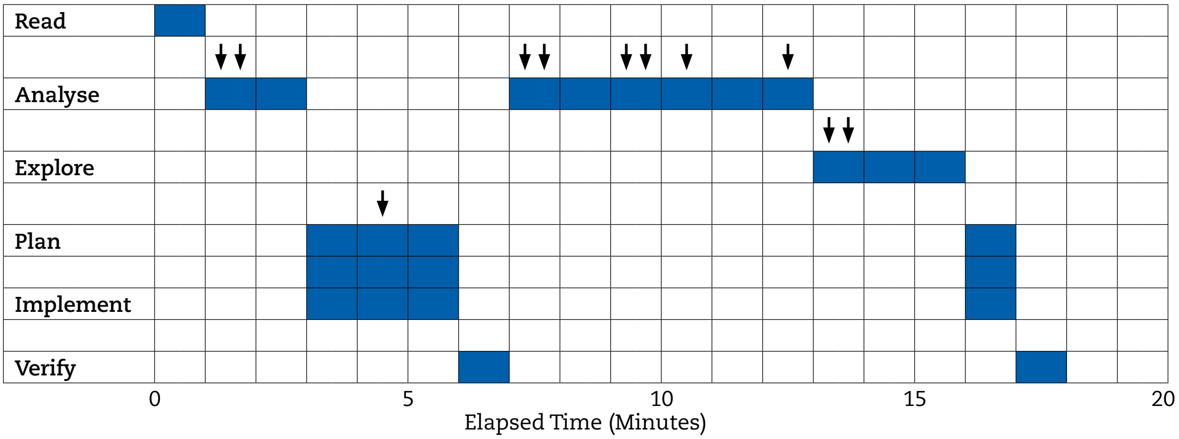


Source: Schoenfeld (1989), "Teaching mathematical thinking and problem solving".

An examination of the students' problem-solving activities shows that the pair of students read the problem for a minute or less. They then spent the rest of their time involved in exploring and working on an approach with no signs of any time being allocated to reflective thought focusing on monitoring progress toward a solution. Schoenfeld indicated that this activity-time graph is representative of 60% of the cases of student work he has observed in uninitiated problem solvers.

Figure 4.3 shows the work of a university mathematics professor when confronted with a difficult two-stage problem. This display stands in stark contrast to that of the uninitiated problem-solvers in Figure 4.2.

Figure 4.3 **A mathematician's problem-solving activities over time**



Source: Schoenfeld, A. H. (1989), "Teaching mathematical thinking and problem solving" in L.B. Resnick and B.L. Klopfer (eds.), Toward the Thinking Curriculum: Current Cognitive Research, Association for Supervision and Curriculum Development, Washington, DC., pp. 83-103.

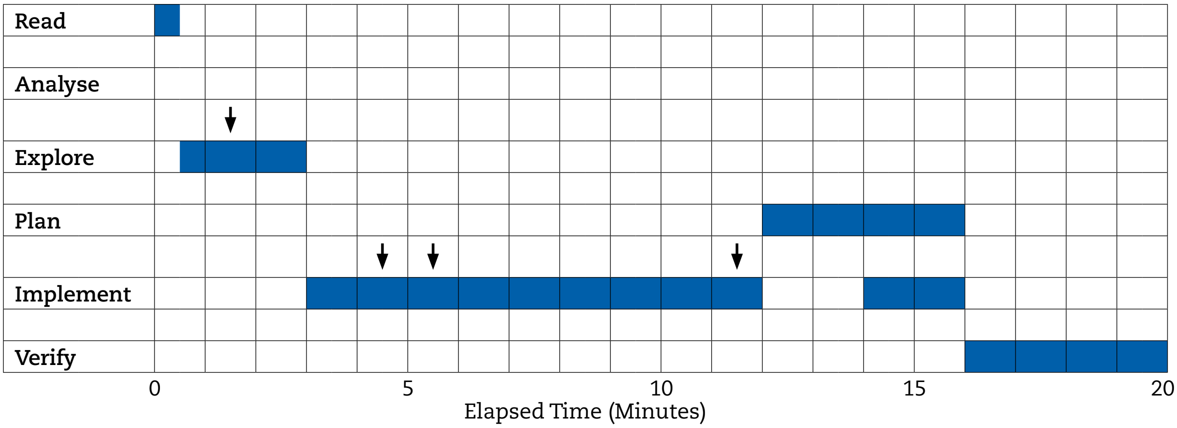
In this graph, the arrows indicate instances where the mathematics professor made an explicit comment about what he was thinking and how it related to the problem solution. Examples were statements such as "Hmm, I don't know exactly where to start here" followed by a shift to two minutes of analysing the problem. This then led to three minutes of cycling through planning and implementation followed by a minute in verification of the first part of the problem. He then returned to the second part of the problem, again making several comments reflective of his metacognitive activities relative to the problem and its solution.

Schoenfeld indicates that he informs students in his class that he will be circulating among the groups while they work on problems. When he visits their group, he is likely to ask one or more of three questions (see Schoenfeld, 1992:356):

- What (exactly) are you doing? (Can you describe it precisely?)
- Why are you doing it? (How does it fit into the solution?)
- How does it help you? (What will you do with the outcome when you obtain it?).

Over the period of a semester, students in the problem-solving class become more adept at answering these questions and even using them as a framework for discussing their solution process among themselves. Figure 4.4 shows an example of an activity-time allocation graph for a pair of students who had completed the problem-solving course while they were working on a problem. Note the change from the activity-time allocation noted in Figure 4.2.

Figure 4.4 **A pair of students' activity-time allocation after a problem-solving course**



Source: Schoenfeld, A. H. (1989), "Teaching mathematical thinking and problem solving" in L.B. Resnick and B.L. Klopfer (eds.), Toward the Thinking Curriculum: Current Cognitive Research, Association for Supervision and Curriculum Development, Washington, DC., pp. 83-103.

The allocations here are much more like that of the experienced mathematician in approaching the problem and its solution. Note the understanding (reading), exploring, planning, implementing with monitoring, cycling back to planning and then back to implementing, with a final segment of time in verification activities. The change, documented in times and activities in the transitions from Figures 4.2 to Figure 4.4, shows the type of observations that evolve in students who are immersed in problem-solving contexts with expert feedback. What is clear from the focus on metacognitive behaviour and the thinking aloud interviews with problem solvers is that self-regulation behaviour as part of problem-solving instruction is an important part of developing mathematical thinking skills.

Beliefs and attitudes and instructional practice

Students' approach and avoidance behaviour is shaped to a great extent by their experiences of learning and their environment (McLeod and Adams, 1989). Students' perseverance on a task is often related to their confidence or self-image relative to their ability to do mathematics and to get "correct" answers. Students have a conception of mathematical problems as "trials" that they must endure and that result in a solution if they are able to remember a particular algorithm. When their first attempts fail, they move on to the next thing in the queue on the test, in the book, or in class.

Results of international comparative studies indicate that attitudes and relating mathematical thinking to games and enjoyable activities result in higher performances on assessments including novel problems and other items eliciting problem-solving skills (Chen and Stevenson, 1995; Randel, Stevenson and Witruk, 2000).

Similarly, teachers' attitudes play a large role in how mathematics is conceptualised, structured and presented in the classroom. When teachers are presented with test-mastery materials alongside problem-solving materials, they often opt to focus on the test-mastery material rather than providing experiences with the heuristics that were also the focus of the lesson. Further, there is the underlying foundation of what mathematics itself is. Is it a static set of facts, concepts, generalisations and structures that can be applied to problems or is it a dynamic study of relationships and models that help bring understanding to structures and the links between them in our ever-changing world? The ways in which students' experiences with mathematical thinking occur in the classroom and elsewhere help provide a significant portion of the beliefs and attitudes they form about problems and their solutions (Dossey, 1992; Cooney, 1985; Mevarech and Kramarski, 2014; Thompson, 1985).

Summary

Research on problem solving, along with models considering stages within the process itself, has provided mathematicians and mathematics educators at all levels with a great deal of information and directions for further research. However, much remains to be done to further the development of learners' capabilities to implement heuristics in profitable ways in their problem-solving work in mathematics classes and in their use of mathematics in the world beyond. It is clear that the development of metacognitive actions shows promise towards developing the conditions where the use of heuristics may be profitable. The solving of real problems in classroom learning and in the practice of mathematics, pure or applied, is still a core issue in our world.

References

Allen, G. (1888). "Genius and talent", *Eclectic Magazine of Foreign Literature, Science, and Art*, Vol. 47/4, pp. 448-458.

Begle, E.G. (1979), *Critical Variables in Mathematics Education Findings from a Survey of the Empirical Literature*, Mathematical Association of America, Washington, DC.

Bransford, J.D., A.L. Brown and R.R. Cocking (eds.) (1999), *How People Learn: Brain, Mind, Experience, and School*, National Academy Press, Washington, DC.

Brown, J.S., A. Collins and P. Duguid (1989). "Situated cognition and the culture of learning", *Educational Researcher*, Vol. 18/1, pp. 32-42.

Chen, C. and H.W. Stevenson (1995), "Motivation and mathematics achievement: A comparative study of Asian-American, Caucasian-American, and east Asian high school students", *Child Development*, Vol. 66/4, pp. 1215-1234.

Charles, R.I. and F.K. Lester (1984), "An evaluation of a process-oriented instructional program in mathematical problem solving in Grades 5 and 7", *Journal for Research in Mathematics Education*, Vol. 15/1, pp. 15-34.

Cooney, T.J. (1985), "A beginning teacher's view of problem solving", *Journal for Research in Mathematics Education*, Vol.16/5, pp. 324-336.

Cooney, T.J. and K.B. Henderson (1971), "A model for organizing knowledge", *Educational Theory*, Vol. 21/1, pp. 50-58.

Dossey, J.A. (1992), "The nature of mathematics: Its role and its influence", in D. Grouws (ed.), *Handbook for Research on Mathematics Teaching and Learning*, Macmillan, New York, pp. 39-48.

Dossey, J.A., I.V.S. Mullis, M.M. Lindquist and D.L. Chambers (1988), *The Mathematics Report Card: Are We Measuring Up? Trends and Achievement Based on the 1986 National Assessment*, Educational Testing Service, Princeton, NJ.

Duncker, K. (1945), "On problem solving", *Psychological Monographs*, Vol. 58/5, Whole No. 270.

Frensch, P.A. and J. Funke (eds.) (1995), *Complex Problem Solving: The European Perspective*, Erlbaum Associates, Hillsdale, NJ.

Funke, J. (2010), "Complex problem solving: A case for complex cognition?", *Cognitive Processing*, Vol. 11/2, pp. 133-142, http://dx.doi.org/10.1007/s10339-009-0345-0.

Garfalo, J. and F.K. Lester (1985), "Metacognition, cognitive monitoring, and mathematical performance", *Journal for Research in Mathematics Education*, Vol. 16/3, pp. 163-176.

Gardiner, A. (2008), "The art of problem solving" in T. Gowers, J. Barrow-Green and I. Leader (eds.), *The Princeton Companion to Mathematics*, Princeton University Press, Princeton, NJ, pp. 955-966.

Gillings, R. J. (1972), *Mathematics in the Time of the Pharaohs*, MIT Press, Cambridge, MA.

Hadamard, J. (1945), *The Psychology of Invention in the Mathematical Field*, Dover Publications, New York.

Henderson, K. B. (1967), "A model for teaching mathematical concepts", *The Mathematics Teacher*, Vol .60/6, pp. 573-577.

Lave, J. and E. Wenger (1991), *Situated Learning: Legitimate Peripheral Participation*, Cambridge University Press, New York.

Lesh, R. (1985), "Conceptual analyses of problem-solving performance" in E.A. Silver (ed.), *Teaching and Learning Mathematical Problem Solving: Multiple Research Perspectives*, Erlbaum Associates, Hillsdale, NJ, pp. 309-330.

Lester, F.K., J. Garofalo and D.L. Kroll (1989), "Self-confidence, interest, beliefs, and metacognition: Key influences on problem-solving behaviour", in D.B. McLeod and V.M. Adams (eds.), *Affect and Mathematical Problem Solving: A New Perspective*, Springer-Verlag, New York, pp. 75-88.

McLeod, D and V. Adams (1989), *Affect and Mathematical Problem Solving: A New Perspective*, Springer-Verlag, New York.

Mayer, R.E. (1990). "Problem solving" in M.W. Eysenck (ed.), *The Blackwell Dictionary of Cognitive Psychology*, Basil Blackwell, Oxford, UK, pp. 284-288.

Mayer, R.E. and M.C. Wittrock (2006), "Problem Solving", in P.A. Alexander and P.H. Winne (eds.), *Handbook of Educational Psychology*, 2nd edition, Erlbaum Associates, Mahwah, NJ:

Mevarech, Z. and B. Kramarski (2014), *Critical Maths for Innovative Societies: The Role of Metacognitive Pedagogies*, Centre for Educational Research and Innovation, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264223561-en.

OECD (2013), *PISA 2012 Assessment and Analytic Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264190511-en.

OECD (2005), *Problem Solving for Tomorrow's World: First Measures of Cross-Curricular Competencies from PISA 2003*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264006430-en.

Poincaré, H. (1952), "Mathematical creation", in B. Ghiselin (ed.), *The Creative Process*, New American Library, New York, pp. 33-42.

Pólya, G. (1945), *How to Solve It: A New Aspect of Mathematical Method*, Princeton University Press, Princeton, NJ.

Polya, G. (1954), *Mathematics and Plausible Reasoning. Vol. I-II,* Princeton University Press, Princeton, NJ.

Randel, B., H.W. Stevenson and E. Witruk (2000), "Attitudes, beliefs, and mathematics achievement of German and Japanese high school students", *International Journal of Behavioral Development*, Vol. 24/2, pp. 190-198.

Rota, G.C. (1986), "More discrete thoughts" in M. Kac, G.C. Rota, and J. T. Schwartz, *Discrete Thoughts: Essays on Mathematics, Science, and Philosophy*, Birkhaüser-Boston, Boston, MA, pp. 263-264.

Ryle, G. (1947), *The Concept of Mind*, Hutchinson, London.

Schoenfeld, A. H. (1992), "Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics" in D. Grouws (ed.), *Handbook for Research on Mathematics Teaching and Learning*, Macmillan, New York, pp. 334-370.

Schoenfeld, A. H. (1989), "Teaching mathematical thinking and problem solving" in L.B. Resnick and B.L. Klopfer (eds.), *Toward the Thinking Curriculum: Current Cognitive Research*, Association for Supervision and Curriculum Development, Washington, DC., pp. 83-103.

Schoenfeld, A.H. (1985), "Metacognitive and epistemological issues in mathematical understanding" in E.A. Silver (ed.), *Teaching and Learning Mathematical Problem Solving: Multiple Research Perspectives*, Erlbaum Associates, Hillsdale, NJ, pp. 361-379.

Silver, E.A. (1987).,"Foundations of cognitive theory and research for mathematics problem-solving instruction" in A. Schoenfeld (ed.), *Cognitive Science and Mathematics Education*, Erlbaum Associates, Hillsdale, NJ, pp. 33–60.

Silver, E.A., N. Branca and V. Adams (1980), "Metacognition: The missing link in problem solving" in R. Karplus (ed.), *Proceedings of the Fourth International Conference for the Psychology of Mathematics Education*, University of California Press, Berkeley, CA, pp. 213-221.

Stanic, G.M.A. and J. Kilpatrick (1989), "Historical perspectives on problem solving in the mathematics curriculum" in R.I. Charles and E.A. Silver (eds.), *Research Agenda for Mathematics Education: The Teaching and Assessing of Problem Solving*, National Council of Teachers of Mathematics, Reston, VA, pp. 1-22.

Steen, L.A. (ed.) (1990), *On the Shoulders of Giants: New Approaches to Numeracy*, National Academy Press, Washington, DC.

Thompson, A. (1985), "Teachers' conceptions of mathematics and the teaching of problem solving" in E.A. Silver (ed.), *Teaching and Learning Mathematical Problem Solving: Multiple Research Perspectives*, Erlbaum Associates, Hillsdale, NJ, pp. 281-294.

Törner, G., A.H. Schoenfeld and K. M. Reiss (eds.) (2007). "Problem solving around the world: Summing up the state of the art", *ZDM*, Vol/5-6, pp. 353-353.

PART II

Dynamic problem solving as a new perspective

Chapter 5

The PISA 2012 assessment of problem solving

By
Dara Ramalingam, Ray Philpot
Australian Council for Educational Research, Melbourne, Australia.
and Barry McCrae
University of Melbourne and Australian Council for Educational Research, Australia.

In 2012, the OECD's Programme for International Student Assessment (PISA) featured a computer-based test of problem-solving competency. Computer delivery enabled the assessment to include for the first time in such a large-scale survey, problems that require the solver to interact with the problem situation to uncover information that is not explicitly disclosed from the outset. These types of problems occur frequently in everyday life when dealing with technological devices such as mobile telephones and vending machines. This chapter presents the PISA 2012 definition of problem solving, outlines the development of the assessment framework and discusses its key organising elements: the problem context, the nature of the problem situation and the four problem-solving process groups. It then presents sample items with commentary.

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

Introduction

The Programme for International Student Assessment (PISA) is a triennial assessment of 15-year-olds conducted by the Organisation for Economic Co-operation and Development (OECD) with the fundamental aim of examining how well-prepared students from different countries are for life beyond school. The first PISA assessment took place in 2000, with 31 participating countries and economies. In 2012 there were 67 participants, comprising 34 OECD countries and 33 partner economies.

Every PISA cycle assesses three core domains: reading, mathematics and science. The resulting estimations of student abilities enable groups of students to be compared between and within PISA-sampled populations. Measures of ability are linked to background variables, such as gender and socio-economic background, enabling researchers to explore which factors are correlated with student performance. For each domain it is assumed that the underlying trait of interest forms a continuum: test items can be placed at a point on the continuum according to how much or how little of the trait is needed to successfully complete the item, and students can be placed at a point on the same continuum according to how much of the underlying trait being assessed they possess. A form of the Rasch model is used to scale the student data and derive the comparative measures reported, and the single scale developed for an assessment domain is used to describe proficiency at different levels of that scale. A detailed description of this model can be found in the *PISA 2000 Technical Report* (Adams and Wu, 2003).

In each PISA cycle it is typical for one or more additional domains to be offered as optional assessments. In PISA 2003, an optional, paper-based assessment of problem solving was available. Forty countries or partner economies participated in this assessment. The problem-solving test represented an attempt to assess a cross-curricular competency within PISA, something that has been declared as “an integral part of PISA” from its inception (OECD, 1999:8).

Developments since the conceptualisation of the 2003 assessment in the areas of complex problem solving, transfer of problem solving skill across domains, computer-based assessment and large-scale assessments of problem-solving competency made the idea of revisiting an assessment of problem solving within PISA an attractive one (e.g. Blech and Funke, 2005; Funke and Frensch, 2007; Greiff and Funke, 2008; Klieme, 2004; Klieme, Leutner and Wirth, 2005; Leutner et al., 2004; Mayer, 2002; Mayer and Wittrock, 2006; O’Neil, 2002; Osman, 2010; Reeff, Zabal and Blech, 2006; Wirth and Klieme, 2004). Accordingly, the decision was made to include a computer-based assessment of problem solving in the PISA 2012 cycle, with the intention of modifying and extending what was accomplished in the PISA 2003 assessment. Most notably, direction was given by the PISA Governing Board to reconsider the definition of problem solving for the assessment, and to consider how the new mode of delivery of the assessment (computer delivery) could best be harnessed to broaden the assessment construct.

The development of a new framework for assessing problem solving and the assessment itself was undertaken by a consortium led by the Australian Council for Educational Research, under the guidance of an eight-member expert group established for this purpose by the OECD. The names and affiliations of this expert group are included as Annex 5.A1. The problem-solving assessment was undertaken by 28 countries and 16 partner economies. The complete list of participants in the 2012 problem solving assessment is given in Annex 5.A2.

Major issues identified

As work on the framework commenced, it became apparent that there were a number of issues that needed clarification to ensure a coherent approach. The most central of these issues were the nature of the relationship between problem solving and other assessment domains, and the place of so-called complex problem solving in the assessment. Each of these issues is outlined below.

The relationship between problem solving and other domains

A fundamental question that must be addressed in attempting to define and operationalise problem solving, is whether there is such a thing as competency in "general problem solving" that is independent of domain-specific knowledge and strategies.

Until the early 1970s, research on problem solving focused on problems that could be easily administered in the laboratory (for example Ewert and Lambert's (1932) "disk" problem, later known as the Tower of Hanoi problem). These kinds of problem are sometimes called "simple" or "analytical" problems. The assumption was that the skills and cognitive processes that were used to solve these problems were generalisable, that is, would be the same as those used to solve more complicated, real-life problems. Perhaps the most influential example of this line of thinking was Newell, Shaw and Simon's (1959) work on the "general problem solver", which assumed problem solving to be a general ability, independent of domain and context.

As research progressed, however, it became apparent that the assumption that skills used in solving analytical problems could be easily transferred to solving real-life problems was not supported. Clear evidence emerged showing the important influence of context: the situation in which a problem is embedded influences the strategies that the solver uses (Kotovsky, Hayes and Simon, 1985; Funke, 1992) and the processes used to solve real-life problems will not necessarily be the same across domains (Sternberg, 1995).

The realisation that skills used in solving problems in the laboratory were not easily generalisable to real-life problems led to a divergence in the focus of research in North America and Europe. Funke and Frensch (2007) describe the evolution of these two schools of research in detail. In summary, with regards to those aspects relevant to the problem-solving framework for PISA 2012, researchers in North America (in particular, Herbert Simon and colleagues – e.g., see Bhaskar and Simon, 1977), convinced that there was no "general" problem-solving ability, began studying the processes involved in problem solving in specific, natural knowledge domains, later moving to exploring whether skills could be transferred between domains (e.g. Mayer and Wittrock, 1996; Mayer, 2002). In contrast, researchers in Europe turned their focus to the consideration of "complex problem solving", using problems designed to be similar to real-life problems. The meaning and relevance of complex problem solving to the PISA 2012 assessment is discussed in detail in the following section.

While these two traditions bring different approaches to the study of problem solving there is agreement in some key aspects of their findings; in particular, research has consistently found that individuals with high levels of problem-solving competency use knowledge and strategies from specific domains (e.g. Mayer, 1992; Funke and Frensch, 2007). Both traditions informed development of the PISA 2012 problem-solving framework.

The centrality of complex problem solving

What is complex problem solving? In what ways is it different to solving analytical problems? Wirth and Klieme (2004) characterise the analytical aspects of problem-solving competence as those required when the goal is well-defined, and all the information needed by the problem solver is available at the outset. Solving the problem requires an evaluation of how to move efficiently move from the given state to the goal state. No exploration or incorporation of feedback is necessary to achieve this. In contrast, complex problem solving includes the need to explore and use the new information learned in solving the problem. Funke (1991) refers to the difference in information available at the outset of a problem as a difference in "transparency" of the problem space.[1] Complex problem-solving tasks arise during the operation of many everyday devices, such as mobile phones, microwave ovens and MP3 players.

There is evidence that complex problem-solving tasks are measuring something different to that assessed in traditional tests of reasoning or intelligence. As part of their PISA 2000 field trial, Germany administered an assessment containing both analytical and complex problems. Students also completed a set of items assessing analogical reasoning, which was used to estimate their intelligence. The resulting data enabled Wirth and Klieme (2004) to explore the relationship between these two kinds of problem solving, and the relationship of each to general intelligence (Chapter 2). Their findings suggest that while analytical problem-solving competence is strongly correlated with intelligence (or more specifically analogical reasoning), complex problem solving is a distinct ability. The finding that complex problem solving is not the same as global intelligence is also generally supported by the work of Funke and Frensch (2007).

Taken together, the importance of complex problem solving in everyday life and the resulting ecological validity gained by including such problems in the assessment, the evidence that complex problem solving measures something different than general intelligence, and the fact that domain-based problem solving is already assessed in PISA's core domains, provided strong justification to make complex problem-solving tasks a major focus of the PISA 2012 assessment. The following sections further explore the framework used to assess this kind of problem, describing its evolution and the key features of the final 2012 framework.

Evolution of the new problem-solving framework

Initial development of the framework began with a review of the research literature and a meeting of experts in mid 2009, and continued over a series of meetings over the next two years. This section discusses significant issues that arose during the process of framework development and the reasoning behind the resolutions reached.

The assessment of complex problem solving

The most important decision resulting from the research review and discussions has already been indicated, namely that the central focus of the assessment should be complex problems that require exploration of the problem space for their solution: the skills involved in solving these problems are increasingly important in everyday life, and the fact that the assessment was to be computer delivered offered the opportunity, for the first time, to include such problems in a large-scale, international assessment.

Furthermore, it was decided that for the purposes of the framework and subsequent test development, the crucial differentiating factor between analytical and complex problems would be the difference in their transparency, i.e. the amount of information disclosed at the outset of the problem. This decision was made because the difference is a central one and is the most readily interpretable by a broad audience. The other features of complex problems identified by Funke and others (e.g. Funke, 1991) might be represented in the items developed, but would not be emphasised in the framework as defining features of complex problems.

Finally, notwithstanding the conclusion that the assessment should focus on complex problems, it was decided that analytical problems should also be included. The justification for this decision was twofold. Firstly, as previously indicated, there is some evidence that knowledge acquisition and its application in solving complex problems are distinct from the typical skills used in solving analytical problems (e.g. Klieme, 2004; Wirth and Klieme, 2004; Leutner and Wirth, 2005). Given this distinction, it was felt that by including both kinds of problems a more comprehensive assessment of problem-solving competency would be gained. Secondly, as insurance, since while several smaller-scale projects had made use of complex problems in assessment, their reliability had never been tested in a large-scale international assessment (although promising results were obtained in the PISA field trial during 2011). It was agreed that a ratio of complex problems to analytical problems of about 2:1 would be appropriate in the main study.

Terminology

A major challenge was finding acceptable terminology to describe the two different types of problems thus far described as "analytical" and "complex". To reiterate, the key difference between these two kinds of problems for the purposes of PISA 2012 is that in the former, all information needed to solve the problem is available to the solver at the outset, while in the latter, the problem situation needs to be explored to uncover additional information needed to solve the problem. The challenge lay in finding a pair of terms to describe these problems that 1) was accurate and consistent with the body of research on problem solving; and 2) effectively communicated the key difference between the two kinds of problems to researchers, policy makers and educators worldwide.

Early versions of the framework used the terms "static" and "dynamic". It was pointed out though, that while "static" seemed an adequate description for analytic problems, the term "dynamic" was not appropriate for complex problems since in the research literature this term is often reserved for describing systems that can change autonomously – that is, the situation has the ability to change without direct action from the problem solver.

Another set of terms considered and rejected was "simple" versus "complex". The reason for rejecting this option was that characterising problems in which all information is available at the outset as "simple" might lead to the conclusion that all problems of this kind are straightforward, offering little challenge to the solver, while problems in which exploration is required are more difficult. In reality, both kinds of problem can be very easy, very difficult, or anything in between. For example, problems in which all information is disclosed may be difficult because a large amount of information needs to be considered, or because sophisticated reasoning skills are required. Conversely problems requiring exploration may be very easy if, for example, the exploration is guided and the relationship between the variables in the system is easy to determine. Using the terminology "simple" and "complex" was therefore judged to be misleading and abandoned as an option.

The pair of terms "intransparent" and "transparent" was also rejected. The term "intransparent" is sometimes used in the research literature to describe problems for which complete information about the problem is not available at the outset. However, feedback from reviewers was that this pair of terms would not be clear or communicative to the broader PISA audience, in part because "intransparent" is not accepted English vocabulary.

The final decision was to use the terms "static" and "interactive" to describe the nature of the corresponding problem situations. The term "interactive" has an everyday meaning that relates well to the key feature of complex problems in PISA 2012 – that interaction is required to discover undisclosed information. It was therefore felt that this term would communicate well to the broad PISA audience and the contrast with "static" – to describe problem situations in which no exploration is needed – is implied. It was recognised, however, that using these terms might convey the misconception that, unlike the complex problems in the assessment, static problems would not make use of the interactive possibilities enabled by computer delivery. It has been important, therefore, to keep communicating that for both interactive and static items, exploiting the opportunities offered by computer delivery would be an integral part of the assessment. Both kinds of items are "interactive" in this sense of the word, since the user must always "interact" with the computer to successfully solve the items.

Problem-solving processes

The PISA construct of problem-solving competency consists of many elements, not all of which can be taken into account and varied under the time constraints and other conditions governing PISA assessments. The most important elements need to be identified so that they can be varied to ensure the construction of an assessment containing items which have an appropriate range of difficulty and which provide a broad coverage of the domain. In addition to the nature of the problem situation (static or interactive), the expert group decided that another major organising element should be problem-solving processes, that is, the cognitive processes involved in solving problems. A review of research from Pólya (1945) onward revealed that, although different authors conceive the cognitive processes involved in problem solving in different ways, there is relatively little deviation from the processes described by Pólya: understanding the problem, devising a plan, executing the plan and reflecting on one's solution.

The common cognitive processes identified have included exploring, understanding, formulating, representing, planning, executing, reflecting, communicating and monitoring (e.g. Baxter and Glaser, 1997; Bransford et al., 1999; Mayer and Wittrock, 1996, 2006; Vosniadou and Ortony, 1989). The challenge faced by the expert group was to group these processes in a way that was both consistent with research and minimised the number of groups, so that each one could be targeted with a sufficient number of items in the assessment. This was achieved in part by reflecting on the stages typically involved in solving the type of problems that the PISA 2012 assessment would focus on, namely those in which undisclosed information needed for their solution had to be discovered. It was decided that the following groupings would constitute the problem-solving processes for PISA 2012: 1) exploring and understanding; 2) representing and formulating; 3) planning and executing; and 4) monitoring and reflecting. Descriptions of these process groups are given in a later section.

The opportunities of computer delivery

As has already been noted, from the outset it was specified that the PISA 2012 assessment of problem solving would be computer delivered. The importance of this decision cannot be overestimated. In particular, it made it possible to include problems that are interactive in nature in the PISA 2012 sense: it is not possible to include items in a paper-and-pencil assessment which require exploration to uncover new information. Had the assessment been delivered on paper, such problems could not have been included despite their obvious primacy in the research literature.

Furthermore, a general advantage of computer delivery is the increased range of response formats that become possible. The PISA 2003 paper-and-pencil assessment of problem solving was restricted to selected-response items, in which an answer is chosen from a number of possibilities (multiple-choice questions being a typical example) and constructed-response items, in which students generate their responses. With computer delivery, these traditional question types remain possible, but answers may be expressed in many other ways, for example, by dragging and dropping an object from one place to another, selecting options from a drop-down menu, or highlighting part of a diagram. This range of possible response formats can be used whether the problem is static or interactive in nature: as mentioned above, although static items do not require exploration to uncover new information, they can be interactive in the sense that they require interaction with the computer interface to provide a solution solution.

Another major advantage of computer delivery is the opportunity it offers for collecting information about what students do as they attempt to solve a problem and the potential for using such information in scoring. In a large-scale paper-and-pencil assessment, the quality of students' thinking can essentially only be judged from their relatively brief final answers. It is not time- or cost-effective to have students include their working and then assess this working. However, computer delivery makes it possible to collect a wealth of process data that may be used to enhance an estimate of a student's problem-solving competency.

Three main kinds of process data can be collected. First, we can collect information about the types of interaction students perform when solving a problem. Did they click on a particular button?

Drag and drop an object from one place to another? Select an item from a drop-down menu? All such interactions (and many others) can be recorded. Second, we can collect information about the frequency and pattern of interactions. Finally, we can collect information about the timing of interactions and so know precisely when each action was performed. When taken together, these three kinds of data can be used to examine the problem-solving strategies a student employed. The final section of this chapter includes examples which illustrate how such process data can be used to contribute to the scoring of an item so that we are assessing more than just the final answer.

Finally, when discussing the opportunities offered by computer delivery, consideration of the relationship of information and communications technology (ICT) to the construct of problem solving becomes an important question: do we include *the use of technology* as part of what is to be assessed? There is precedent for taking such an approach – for example, in the OECD's Survey of Adult Skills conducted as part of the Programme for the International Assessment of Adult Competencies (PIAAC). PIAAC is a face-to-face sample household survey administered to adults aged between 16 and 65. Together with reading component skills, literacy and numeracy, the 2011 administration of this survey included an assessment of problem solving in technology-rich environments. In PIAAC there is no definition of problem solving in and of itself: the definition in the framework is of "problem solving in technology-rich environments" and explicitly includes the use of technology (PIAAC Expert Group in Problem Solving in Technology-Rich Environments, 2009:9).[1]

The focus of the problem-solving assessment in PIAAC is thus on "information-rich" problems, and solving each problem relies on the use of one or more computer software applications. This constitutes a major difference from the PISA 2012 assessment of problem solving, in which although the items are designed to exploit the possibilities of computer delivery, the use of *software applications* is not assumed to be part of the definition of problem solving and is therefore not assessed. A practical consequence of this is that only foundational ICT skills (such as using a mouse and keyboard) are assumed for the PISA assessment. The focus of PISA is on the cognitive processes used to solve a problem rather than on the use of software applications to do so.

The key features of the PISA 2012 problem-solving framework

The definition of problem-solving competency for PISA 2012 is as follows:

> **Problem-solving competency** is an individual's capacity to engage in cognitive processing to understand and resolve problem situations where a method of solution is not immediately obvious. It includes the willingness to engage with such situations in order to achieve one's potential as a constructive and reflective citizen. (OECD, 2013:122)

A full discussion of all features of the definition is given in the framework (OECD, 2013); only the main features are highlighted here, selected for their relevance to this chapter's content. Note that the term "competency" is used to mean far more than the application of knowledge and skills and is intended to encompass a range of psychosocial resources available to an individual (OECD, 2003). The second sentence of the definition highlights that the use of knowledge and skills to solve a problem is affected by motivational and affective factors (e.g. Mayer, 1998).

There are three key organising elements in the PISA 2012 framework. The first of these is the problem context, which is classified in two ways: as either technology or non-technology and as social or personal. It was intended that representing a wide range of contexts would both ensure that the tasks were generally of interest to 15-year-olds and control for prior knowledge overall, so that no one group of students would be consistently advantaged or disadvantaged.

Problems classified as having a technology context are based on the functioning of a technological device. In a typical unit of this kind, students explore the functionality of a device prior to using the knowledge gained to carry out an action. Other problems, such as task scheduling or route planning, are classified as having a non-technology context. Problems classified as personal are those in which the setting relates to the self, family or peer groups. Social problems are set in the community or society more broadly.

The second organising element, which has already been discussed at length, is the nature of the problem situation.

The final organising element in the 2012 framework is the problem-solving process. Each item is categorised as principally requiring one of the following four processes:

1) **Exploring and understanding** involves both exploring (interacting with the problem space to find undisclosed information) and demonstrating an understanding of the information discovered in the process of exploration. Ultimately, the aim of exploration is to build a mental representation of the information present in the problem, not as a whole, but in terms of constituent elements.

2) The aim of **representing and formulating** is to build a mental model of the problem situation as a whole. This might involve constructing a representation of the problem (in tabular, graphical, symbolic or verbal form) and formulating hypotheses by considering the relationships between the different elements of the problem.

3) **Planning and executing** includes both planning how to solve a problem (for example by setting subgoals) and carrying out the plan. In recognition of the importance of being able to carry through a solution to a successful conclusion, the number of items in the assessment targeting this process was greater than for the other three processes.

4) **Monitoring and reflecting** includes tracking one's progress toward a goal and taking action to divert course where necessary, and reflecting on and communicating solutions from different perspectives. In recognition of the fact that monitoring and reflecting are likely to take place throughout the process of reaching a solution and that it is therefore assessed indirectly in items targeting other processes, fewer items specifically targeted this process than targetted the other three processes.

It is not assumed that all of these processes are necessary for solving a particular problem, nor that they are sequential: when solving authentic problems, individuals may well need to move beyond a linear, step-by-step model. In practice it is often necessary to cycle through these processes as progress is made towards a solution.

Developing the consistency

This section describes the structure of the PISA 2012 problem-solving assessment and the opportunities and challenges that arose in the process of task development.

The structure of the assessment

Countries participating in the assessment of problem solving were required to administer it on the same day as the other PISA survey components. This meant that each student undertaking the problem solving assessment had to first complete the two-hour test of maths, science and reading, followed by a half-hour background questionnaire. The problem-solving assessment itself was relatively short, with each student completing 40 minutes of assessment material. The material presented was rotated: the total pool contained 80 minutes of material : spread across four 20-minute clusters.. The background questionnaire completed by all students contained two sets of items directly relevant to problem solving: 1) a set of items measuring some motivational and affective factors related to problem solving competency; and 2) vignettes (situational judgement tasks) designed to gather information on students' problem-solving strategies in different scenarios (OECD, 2013).

Devising complex (interactive) problems

Devising complex (interactive) problems, the emphasis was on everyday problem situations that often arise when interacting with an unfamiliar device (such as a ticket vending machine, air-conditioning system or mobile phone) for the first time. Some of these devices, such as vending machines, were modelled as finite state machines, that is, systems with a finite number of states, input signals and output signals. The system's next state is determined by its current state and the specific input signal. Other problem situations, such as controlling an air conditioner, involve numerical input and output variables that are related in some way. These situations were implemented as systems of linear structural equations (see Chapter 6). Due to testing time constraints, it was not possible to include the kinds of detailed simulated worlds used by many researchers, starting with Dörner (e.g. Dörner, 1975).

Task development

While it is true that an assessment framework must guide test development, in a large and multifaceted project with as short a timeline as PISA, framework development and test development must necessarily occur in parallel. This, however, has the advantage of allowing each process to inform the other, with test developers and expert group members engaging in combined discussion. The presence of reciprocal influences between framework and test development was a strong feature during work on the PISA 2012 assessment of problem solving.

In early item development, the test developers deliberately developed a wide range of tasks in order both to understand more clearly the expert group's conceptualisation of problem solving for PISA 2012 and to reveal to the expert group where the framework needed enhancement and refinement. For example, one early task involved understanding patterns, in particular, how rotating smaller patterned squares that made up a larger shape would change the pattern seen in the overall design. This task was rejected on the grounds that it too closely resembled a traditional intelligence test task, and therefore did not represent the construct that the expert group intended to be measured. Other early tasks were rejected because it was felt that they could legitimately have been included in one of the other PISA assessment domains (reading, mathematics or science). A unit focusing on the rules underlying a game was deemed too close to a PISA mathematical literacy task, while a unit that involved judging the trajectory a ball would follow on a snooker/billiards table with uneven legs might, it was felt, have been included as either a scientific literacy or mathematical literacy task. Finally, a unit centred on reading a set of public transport timetables could have been included as a reading literacy task. Through this process of review of a wide range of items, the expert group and test developers reached a mutual understanding of the boundaries of the problem-solving construct for the purposes of PISA.

The development of instruments for large-scale international assessments comes with inherent challenges. With 44 participating countries and partner economies completing the PISA 2012 problem-solving assessment, and 52 different national versions required to cover all language-and-country combinations, the most obvious challenge lay in constructing material that would not consistently offer an advantage to any one country or language group over another. One strategy used to meet this challenge was to ensure that personnel from the greatest possible number of backgrounds were involved in both developing and reviewing test material. The material for the PISA 2012 problem-solving assessment came from two sources: test development centres and national submissions. Four different test development centres, based in Australia, Germany, Norway and the United States were involved, and all countries participating in PISA had the opportunity to submit material for possible inclusion in the assessment.

In addition, all material approved by the expert group, who particularly looked at the match of material to the framework, was made available to every participating country for review. For every item, reviewers were asked to consider and comment on a wide variety of issues, including the perceived relevance, interest and authenticity of the task for students in their country, whether any cultural concerns were evident, whether any difficulty in translating the item could be anticipated, and, where relevant, whether the coding (scoring) instructions for the item would lead to inconsistency in coding across countries. When the feedback was reviewed, any items in which cultural concerns looked likely to be an issue (or any items that were not generally approved of for any reason) were excluded from further consideration.

A further strategy to ensure that no country, language group or gender was consistently advantaged was to examine the items for differential item functioning (DIF) across groups. Briefly, an item shows DIF if people in different groups who have the same underlying ability on the trait being measured (in this case, problem-solving competency) have a different probability of gaining a certain score on an item. The presence of DIF indicates that an item may be behaving differently across that group of people. The field trial data for all items were examined to identify items that were problematic in this regard and so could not be considered for inclusion in the main survey.

Related to, but distinct from the issue of cultural concerns, are translation issues. The PISA 2012 problem-solving material was translated into 30 different languages and so ensuring cross-language comparability was imperative. Rigorous processes were in place to address this potential issue. At the commencement of item development all test developers were thoroughly briefed on anticipating (and avoiding) potential translation difficulties in item writing. Over several PISA cycles it has been possible to identify many syntactic and vocabulary issues that will cause translation problems. Informed test developers are able to avoid such issues from the earliest stages of item development. In addition, a rigorous process is in place for ensuring the quality and comparability of all translated versions of the test material. Two source versions (one in English and one in French) are provided to countries, which then follow a process of double translation and reconciliation to prepare their version (McCrae, 2009).

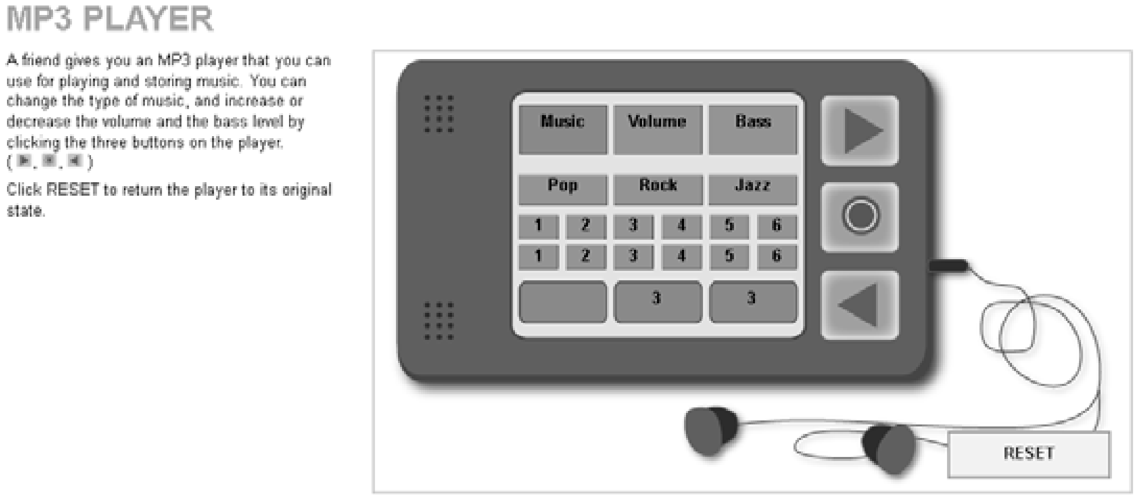
Linked to the issue of ensuring that no country or language group was consistently advantaged in the assessment was the importance of preparing coding guides for constructed-response items. These set out guidelines designed to ensure coding consistency across participating countries. The coding guides included numerous example responses with explanations of the code given. Marginal examples were deliberately included so that coders could gain a sense of exactly where the boundary between codes lay. Representatives of all participating countries were trained in the use of these guides so that they, in turn, could train coders in their own country. In addition, during the coding period countries had access to a central coder query service query service to which they could refer problematic student responses for a ruling. Together, these measures helped to achieve a high level of consistency in coding across countries.

A final challenge in task development for the PISA 2012 problem-solving assessment arose from the requirements to construct problems with which students were not already familiar and that did not need a high level of domain-based knowledge for their solution. Minimising the involvement of domain knowledge ensures that the focus is on measuring the cognitive processes involved in problem solving. This approach constitutes a major difference from the assessment of the other core domains in PISA (reading, mathematics and science): in each of these, the assessments are constructed so that a high level of knowledge in the domain *is* required. For the assessment of problem solving, low-verbal and non-verbal information was used wherever possible in describing problems, and only a basic level of mathematical and scientific knowledge was involved. In reality, ensuring that problems are equally unfamiliar to students is impossible at the individual level: the aim was to achieve this across countries by presenting a variety of contexts so that no one group was consistently advantaged or disadvantaged in this way.

Examples of test material

This section describes the four items from the unit MP3 Player, which was included in the PISA 2012 field trial but was not included in the main survey. Images of the stimulus and item information are taken from the *PISA 2012 Assessment and Analytical Framework* (OECD, 2013).

Figure 5.1. **MP3 player: Stimulus information**



MP3 PLAYER

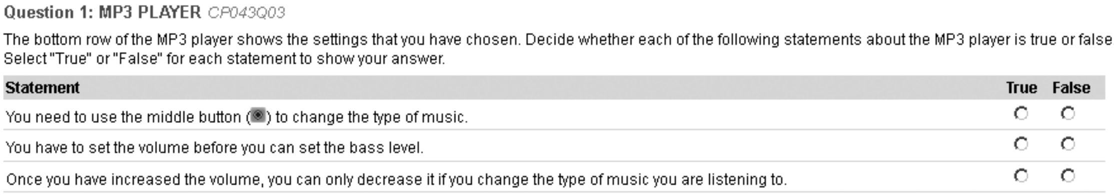
A friend gives you an MP3 player that you can use for playing and storing music. You can change the type of music, and increase or decrease the volume and the bass level by clicking the three buttons on the player. (▶, ◉, ◀)

Click RESET to return the player to its original state.

Source: OECD (2013), PISA 2012 Assessment and Analytical Framework, http://dx.doi.org/10.1787/9789264190511-en.

The premise of the unit MP3 Player (Figure 5.1) is that students have been given an unfamiliar MP3 player which they must find out how to work. Exploration is required to determine the rules governing the device, so the nature of the problem situation is interactive. The problem context is defined as “technology” and “personal”, since it involves the personal use of a technological device.

Figure 5.2. **MP3 player: Item 1**



Question 1: MP3 PLAYER *CP043Q03*

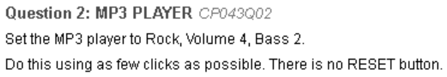
The bottom row of the MP3 player shows the settings that you have chosen. Decide whether each of the following statements about the MP3 player is true or false. Select "True" or "False" for each statement to show your answer.

Statement	True	False
You need to use the middle button (◉) to change the type of music.	○	○
You have to set the volume before you can set the bass level.	○	○
Once you have increased the volume, you can only decrease it if you change the type of music you are listening to.	○	○

Source: OECD (2013), *PISA 2012 Assessment and Analytical Framework*, http://dx.doi.org/10.1787/9789264190511-en.

The first item of the unit (Figure 5.2) presents students with a series of statements about the way the device works. They must identify whether each statement is true or false. The statements were intended to encourage and scaffold initial exploration of the working of the device: the problem-solving process for this item is exploring and understanding. A “reset” button is available so that students can return the MP3 player to its initial state at any time, and their use of this button is unrestricted.

Figure 5.3. **MP3 player: Item 2**



Question 2: MP3 PLAYER *CP043Q02*

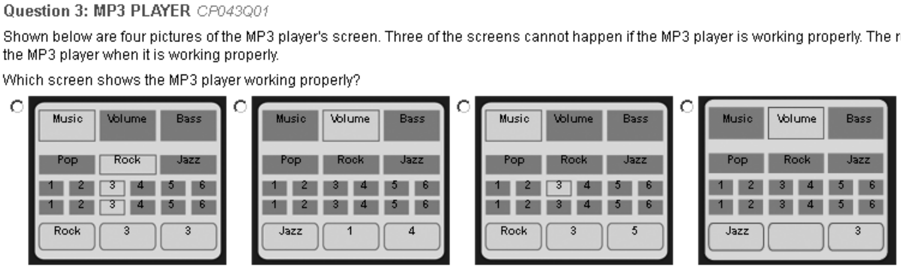
Set the MP3 player to Rock, Volume 4, Bass 2.

Do this using as few clicks as possible. There is no RESET button.

Source: OECD (2013), *PISA 2012 Assessment and Analytical Framework*, http://dx.doi.org/10.1787/9789264190511-en.

In the second item of this unit (Figure 5.3) students must use their knowledge of the device to achieve a given goal. The problem-solving process is planning and executing. The students are explicitly instructed to “Do this using as few clicks as possible”, and efficiency is built into the scoring for this item: data logs show how many steps students take to reach the desired goal state. Students who reach the desired goal state, but do not do this efficiently (i.e. take a large number of steps) receive partial credit for the item, whereas students who reach the goal state efficiently receive full credit for the item. This item is an example of how the PISA 2012 problem-solving assessment made use of the process data generated by students in completing the item, in addition to the final state of the MP3 Player at the end of the item, for scoring student responses.

Figure 5.4. **MP3 player: Item 3**



Question 3: MP3 PLAYER CP043Q01

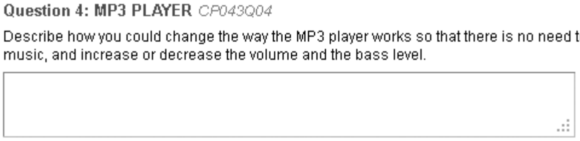
Shown below are four pictures of the MP3 player's screen. Three of the screens cannot happen if the MP3 player is working properly. The remaining screen shows the MP3 player when it is working properly.

Which screen shows the MP3 player working properly?

Source: OECD (2013), *PISA 2012 Assessment and Analytical Framework*, http://dx.doi.org/10.1787/9789264190511-en.

The third item of the unit (Figure 5.4) is a multiple-choice question in which students must identify which of a series of options depicts a state that is possible for the MP3 player. To answer this question correctly requires an overall mental representation of the way the system works, so the problem-solving process is representing and formulating.

Figure 5.5. **MP3 player: Item 4**



Question 4: MP3 PLAYER CP043Q04

Describe how you could change the way the MP3 player works so that there is no need to have the bottom button (◀). You must still be able to change the type of music, and increase or decrease the volume and the bass level.

The final item in the unit (Figure 5.5) asks students to consider how the design of the MP3 player could be changed so that only one arrow button is required to operate the device (rather than two), while retaining all the functionality of the original device. Unlike the earlier items in the unit, this item was not scored automatically: it needed coding by country experts using an online coding system. Credit was given to any answer that described a plausible reconceptualisation of the player. There is no single correct answer: creative thinking can lead to a number of valid responses. The item is classified as monitoring and reflecting.

Conclusion

The development of the framework for the PISA 2012 assessment of problem solving drew together rich research traditions that arose from the notion that domain knowledge plays an important role in problem solving and that problem solving-competence is therefore not independent of domain or context. A clear implication of a review of research was that the main focus of the assessment should be problems in which exploration is required to uncover undisclosed information (classified as interactive problems). These problems are common in everyday life, and solving such problems requires additional skills to those used to solve problems in which all information is available at the outset (static problems).

The other major organising features of the framework are context (items are classified as either technology or non-technology, and social or personal), and problem-solving process, with each item targeting one of four process groups: exploring and understanding, representing and formulating, planning and executing, and monitoring and reflecting.

The PISA 2012 assessment of problem solving moved well beyond what had previously been accomplished in large-scale international assessments. Computer delivery made possible the inclusion of everyday interactive problems, and these were a major focus of the assessment. In addition, computer delivery allowed process data, captured in the course of a student's interactions with a task, to contribute explicitly to scoring, representing another major innovation.

Notes

1 In addition to transparency, Funke (1991) describes five other features of complex problems: 1) polytely; 2) complexity of the situation; 3) connectivity of variables; 4) dynamic developments; and 5) time-delayed effects.

2 "Problem solving in technology rich environments involves using digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks" (PIAAC Expert Group in Problem Solving in Technology-Rich Environments, 2009:9).

References

Adams, R.J. and M.L. Wu (eds.) (2003), *Programme for International Student Assessment (PISA): PISA 2000 Technical Report*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264199521-en.

Baxter, G.P., and R. Glaser (1997), "An approach to analysing the cognitive complexity of science performance assessments", *CSE Technical Report*, No. 452, National Center for Research on Evaluation, Standards and Student Testing (CRESST), Los Angeles.

Bhaskar, R. and H.A. Simon (1977), "Problem solving in semantically rich domains: An example from engineering thermodynamics", *Cognitive Science*, Vol. 1/2, pp. 193-215.

Blech, C. and J. Funke (2005), *Dynamis Review: An Overview About Applications of the Dynamis Approach in Cognitive Psychology*, Deutsches Institut für Erwachsenenbildung, Bonn, www.die-bonn.de/esprid/dokumente/doc-2005/blech05_01.pdf.

Bransford, J.D., A.L. Brown and R.R. Cocking (eds.) (1999), *How People Learn: Brain, Mind, Experience, and School*, National Academy Press, Washington, DC.

Dörner, D. (1975). "Wie Menschen eine Welt verbessern wollten" (How people wanted to improve the world), *Bild der Wissenschaft*, Vol. 12/2, pp. 48-53.

Ewert, P.H. and J.F. Lambert (1932), "Part II: The effect of verbal instructions upon the formation of a concept", *Journal of General Psychology*, Vol. 6/2, pp. 400-413.

Funke, J. (1992), "Dealing with dynamic systems: Research strategy, diagnostic approach and experimental results", *The German Journal of Psychology*, Vol. 16/1, pp. 24-43.

Funke, J. (1991), "Solving complex problems: Human identification and control of complex systems", in R.J. Sternberg and P.A. Frensch (eds.), *Complex Problem Solving: Principles and Mechanisms*, Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 185-222.

Funke, J. and P.A. Frensch (2007), "Complex problem solving: The European perspective – 10 years after", in D.H. Jonassen (ed.), *Learning to Solve Complex Scientific Problems*, Lawrence Erlbaum, New York, pp. 25-47.

Greiff, S. and J. Funke (2008), "Indikatoren der Problemlöseleistung: Sinn und Unsinn verschiedener Berechnungsvorschriften" (Measuring complex problem solving: The MicroDYN approach), in *Bericht aus dem MirdoDYN Projekt*, Psychologisches Institut, Heidelberg.

Klieme, E. (2004). "Assessment of cross-curricular problem-solving competencies", in J.H. Moskowitz and M. Stephens (eds.), *Comparing Learning Outcomes: International Assessments and Education Policy*, Routledge Falmer, London, pp. 81-107.

Klieme, E., D. Leutner and J. Wirth (2005), *Problemlösekompetenz von Schülerinnen und Schülern: Diagnostische Ansätze, theoretische Grundlagen und empirische Befunde der deutschen PISA 2000 Studie (Problem-solving Competency of Students: Assessment Approaches, Theoretical Basics, and Empirical Results of the German PISA 2000 Study)*, VS Verlag für Sozialwissenschaften, Wiesbaden, Germany.

Kotovsky, K., J.R. Hayes and H.A. Simon (1985), "Why are some problems hard? Evidence from the Tower of Hanoi", *Cognitive Psychology*, Vol. 17/2, pp. 248-294.

Leutner, D.E., E. Klieme, K. Meyer and J. Wirth (2004), "Problemlösen" (Problem solving), in M. Prenzel et al. (eds.), *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs*, PISA-Konsortium Deutschland, Waxmann, Münster, Germany, pp. 147–175.

Leutner, D. and J. Wirth (2005), "What we have learned from PISA so far: A German educational psychology point of view", *KEDI Journal of Educational Policy*, Vol. 2/2, pp. 39-56.

Mayer, R.E. (2002), "A taxonomy for computer-based assessment of problem solving", *Computers in Human Behavior*, Vol. 18/6, pp. 623-632.

Mayer, R.E. (1998), "Cognitive, metacognitive, and motivational aspects of problem solving", *Instructional Science*, Vol. 26/1, pp. 49-63.

Mayer, R.E. (1992), *Thinking, Problem Solving, Cognition*, 2nd edition, WH Freeman, New York.

Mayer, R.E and M.C. Wittrock (2006). "Problem solving", in P.A. Alexander and P.H. Winne (eds.), *Handbook of Educational Psychology,* 2nd edition, Lawrence Erlbaum Associates, Mahwah, NJ.

Mayer, R.E. and M.C. Wittrock (1996). "Problem-solving transfer", in R. Calfee and R. Berliner (eds.), *Handbook of Educational Psychology*, Macmillan, New York, pp. 47-62.

McCrae, B.J. (2009). "PISA 2006 test development and design", in R.W. Bybee and B.J. McCrae (eds.), *PISA Science 2006: Implications for Science Teachers and Teaching*, National Science Teachers Association, Arlington, VA, pp. 27-38.

Newell, A., J.C. Shaw and H.A. Simon (1959), "A general problem-solving program for a computer", *Computers and Automation*, Vol. 8, pp. 10-16.

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

OECD (2003), *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264101739-en.

OECD (1999), *Measuring Student Knowledge and Skill: A New Framework for Assessment*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264173125-en.

O'Neil, H.F. (2002), "Perspectives on computer-based assessment of problem solving", *Computers in Human Behavior*, Vol. 18, pp. 605-607.

Osman M. (2010), "Controlling uncertainty: A review of human behavior in complex dynamic environments", *Psychological Bulletin*, Vol. 136/1, pp. 65-86.

PIAAC Expert Group in Problem Solving in Technology-Rich Environments (2009), "PIAAC problem solving in technology-rich environments: A conceptual framework", *OECD Education Working Papers*, No. 36, OECD Publishing, Paris, http://dx.doi.org/10.1787/220262483674.

Pólya, G. (1945), *How to Solve It*, Princeton University Press, Princeton, NJ.

Reeff, J.P., A. Zabal and C. Blech (2006), *The Assessment of Problem-Solving Competencies: A Draft Version of a General Framework*, Deutsches Institut für Erwachsenenbildung, Bonn, . www.die-bonn.de/esprid/dokumente/doc-2006/reeff06_01.pdf.

Sternberg, R.J. (1995), "Expertise in complex problem solving: A comparison of alternative conceptions", in P.A. Frensch and J. Funke (eds.), *Complex Problem Solving: The European Perspective*, Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 295-321.

Vosniadou, S. and A. Ortony (1989), *Similarity and Analogical Reasoning*, Cambridge University Press, New York.

Wirth, J. and E. Klieme (2004), "Computer-based assessment of problem solving competence", *Assessment in Education: Principles, Policy and Practice*, Vol. 10/3, pp. 329-345.

Annex 5.A1. The PISA 2012 Problem Solving Expert Group

PEG Member	Affiliation
Joachim Funke (Chair)	Department of Psychology, Heidelberg University, Germany
Benő Csapó	Institute of Education, University of Szeged, Hungary
John Dossey	Department of Mathematical Sciences, Illinois State University, United States of America
Art Graesser	Psychology, Computer Science, and Institute for Intelligent Systems, The University of Memphis, United States of America
Detlev Leutner	Department of Instructional Psychology, Duisburg-Essen University, Germany
Romain Martin	Université de Luxembourg, FLSHASE, Luxembourg
Richard Mayer	Department of Psychology, University of California, Santa Barbara, United States of America
Ming Ming Tan	Sciences, Curriculum Planning and Development Division, Ministry of Education, Singapore

Annex 5.A2. Countries and partner economies participating in the OECD PISA 2012 problem-solving assessment

- Australia
- Austria
- Belgium
- Brazil
- Bulgaria
- Canada
- Chile
- Colombia
- Croatia
- Cyprus*
- Czech Republic
- Denmark
- Estonia
- Finland
- France
- Germany
- Hong Kong (China)
- Japan
- Hungary
- Ireland
- Israel
- Italy
- Korea
- Macao (China)
- Malaysia
- Montenegro
- Netherlands
- Norway
- Poland
- Portugal
- Russian Federation
- Serbia
- Shanghai – China
- Sweden
- Singapore
- Slovak Republic
- Slovenia
- Spain
- Chinese Taipei
- Turkey
- United Arab Emirates
- England, United Kingdom
- United States
- Uruguay

* Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

Chapter 6

Interactive problem solving: Exploring the potential of minimal complex systems

by

Samuel Greiff

ECCS unit, University of Luxembourg, Luxembourg.

and Joachim Funke

Department of Psychology, Heidelberg University, Germany.

Interactive problem solving (IPS) is considered an increasingly important skill in professional and everyday life as it mirrors the interaction of a human user with dynamic technical and non-technical devices. Here, IPS is defined as the ability to identify the unknown structure of artefacts in dynamic, mostly technology-rich, environments to reach certain goals. Two overarching processes, the acquisition of knowledge and its application, can be theoretically distinguished and this chapter presents two measurement approaches assessing these processes: MicroDYN and MicroFIN, both of which rely on the idea of minimal complex systems (MICS). Whereas MicroDYN models quantitative connections between elements of a problem (i.e. more and less), MicroFIN models qualitative connections between them (e.g. on/off or white/black). This chapter summarises research on these approaches and discusses the implications for educational assessment.

The authors thank the German Research Foundation (DFG Fu 173/13 and Fu 173/14) within the Priority Programme "Competence Models for Assessing Individual Learning Outcomes and Evaluating Educational Processes" (SPP 1293) and the Fonds Nationale de la Recherche Luxembourg (ATTRACT "ASKI21") for funding this work.

Please note that some small parts of this chapter have already been published at psychologie.uni-heidelberg.de/ae/allg/forschun/dfg_komp/microdyn.html.

Introduction

Complexity is an intriguing feature of modern times but one that is difficult to grasp. One of its sources is the increasing impact and the augmented presence of interactive devices in our daily environment. Devices that change dynamically with user interaction have arguably led to a greater amount of interactivity and complexity in today's world. This happens not only in light of newly emerging software tools, which make continuous learning necessary, but also in light of specialised hardware that confronts us with complex interactions: mobile phones, ticket machines, electronic room access keys, copiers and even washing machines now require sequences of interactions to set up these devices and to let them run. To master them, one has to press buttons or use sliders and other continuous controls (such as the accelerator of a car, the volume of an MP3 player or the temperature of a heating system) or even a combination of both. Even in non-technical contexts, humans have to face interactive situations, even though this area is only slightly touched on by research on problem solving – although the Programme for International Student Assessment (PISA) 2015 has made efforts to assess collaborative problem solving. For instance, social interaction may be described as a complex and – through verbal and non-verbal interaction – dynamically changing problem. In a conversation, both continuous changes, such as mood increases or decreases, and discrete changes, such as eye contact or avoidance of it, may occur. In fact, humans experience many problems when interacting with different classes of devices. That is one of the reasons why research on problem solving as an individual ability is of increasing importance and assessment of problem solving skills comes into play in various practical and scientific fields.

When interacting with such devices, the problem solver essentially needs to master two tasks. The first is to understand the device itself, the way it dictates interactions with the user, and the individual skills and abilities required to interact with the device (Dörner, 1986; Funke, 2001). . In general, what a human user has to do in dealing with a complex and interactive device is straightforward: find out how the device works and store this information (knowledge acquisition: find a strategy to build up knowledge and then represent it internally; Mayer and Wittrock, 2006). The second task for the solver is to try to reach the goal state (knowledge application: apply the acquired knowledge to reach a certain goal; Novick and Bassok, 2005). Therefore, these two tasks of knowledge acquisition and knowledge application are of primary interest to understand and assess problem solving. Both are essential in the context of problem-solving research (Funke, 2001; Novick and Bassok, 2005; OECD, 2014; Wüstenberg, Greiff and Funke, 2012).

An interactive problem is one where knowledge acquisition and knowledge application are used to apply non-routine actions to reach a certain goal state (Greiff, Holt and Funke, 2013). Dörner (1996) used an even more detailed approach towards describing interactive problems: according to him, solving a problem demands a series of operations which can be characterised as follows: elements relevant to the solution process are large and manifold (complexity), highly interconnected (connectivity), and changing over time (dynamics). Neither structure nor dynamics are disclosed to the problem solver (intransparency). Finally, the goal structure is not as straightforward as suggested above: in dealing with a complex problem a person is confronted with a number of different subgoals, which are to be weighted and co-ordinated against each other – a polytelic situation.

Each of these attributes of an interactive problem corresponds to one of five demands placed on the problem solver as defined by Dörner (1986): 1) complexity requires reduction of information; 2) connectivity requires model building; 3) dynamics requires forecasting; 4) intransparency requires information retrieval; and 5) polytely requires the evaluation of multiple criteria.

Whereas all five requirements play an important role in interactive problem solving, the first three can be understood as specific dimensions of knowledge acquisition and the last two are subsumed under the overarching process of knowledge application (Fischer, Greiff and Funke, 2012). An assessment needs to either focus on the two overarching processes or on each of the five subdimensions, but some conceptual background is needed in addition to psychometric demands dictated by scientific standards of assessment (e.g. Funke and Frensch, 2007).

In this chapter, we focus our understanding of problem solving mainly on the interaction of a human user with a dynamic and complex device and the corresponding processes of knowledge acquisition and knowledge application (Fischer et al., 2012; Novick and Bassok, 2005). To label this focus adequately, we propose to use the term interactive problem solving (IPS),[1] which emphasises the characteristic interaction between problem solver and problem and the inherent changes of the problem situation. This term is also used in PISA 2012 (OECD, 2014). In this chapter, after an explanation of the concept in the first section, the second section illustrates how to measure IPS in a psychometrically sound way based on the idea of interacting with several complex systems. The third section presents a review of recent studies based on this concept. Finally, in the discussion we present three open questions and provide tentative answers to them.

Interactive problem solving

What is IPS and how can it be defined? The 21st-century offers the citizens and workers of industrialised countries a technology-rich environment. It is taken for granted that a person is able to use mobile communication, household devices, electronically equipped cars, public transportation and various technologies in the workplace, in short that they can use technologies offering a broad range of applications, which are rapidly changing and which often introduce new features and services to their users. These features and services require interaction with mostly automated environments (such as using credit cards for public transport or paying via phone) that are in principle similar to each other but at the same time context-specific and change from region to region. For example, when travelling around the world, one experiences a lot of different systems to make payments for public transport even if in principle the systems are fundamentally the same (Funke, 2001; Greiff and Funke, 2009). Thus, in concentrating on IPS we focus on a 21st-century skill that is not assessable by traditional paper-and-pencil tests as these are hardly interactive – we focus on a competence that can only be assessed by a person interacting with a dynamically reacting environment in a computer-based assessment setting.

From this observation, IPS is defined as the ability to explore and identify the structure of (mostly technical) devices in dynamic environments by means of interacting and to reach specific goals (compare with OECD, 2014). It is worth commenting on the different components of this definition in turn. First, the term "ability" refers to the fact that IPS develops over a lifetime and that it can be learned and fostered, for instance by specific training or by schooling (Frischkorn, Greiff, and Wüstenberg, 2014; Greiff et al., 2013). Second, the two tasks mentioned ("to identify the unknown structure" and "to reach specific goals") refer to the abstract requirements of 1) system identification and exploration (knowledge acquisition); and 2) system control and steering (knowledge application; Funke, 1993). System identification simply means finding out how a given device such as a new mobile phone works and how it reacts to certain inputs. It is usually preceded or accompanied by a phase of exploration, without any specific targets (Kröner, Plass and Leutner, 2005; Wüstenberg et al., 2012). System control simply means to know how to get what you want, for instance making a phone call with the new mobile phone. This control heavily relies on the knowledge acquired beforehand (Funke, 2001; Greiff, 2012). Third, the notion of an "unknown structure" points to the fact that IPS deals with new situations for which a routine solution is not at hand and that even though prior knowledge may play a role, it is neither essential nor necessary to solve the problem (Greiff, Wüstenberg and Funke, 2012). Fourth, the allusion to "(mostly technical) devices in dynamic environments" refers to an important aspect of this type of problem, namely the interaction between a user and the device. This interaction implies that during the course of interaction the state of the device changes, either depending on the type of intervention or dynamically by itself (Buchner, 1995). The resulting dynamics are in contrast to simple problems that are static in nature such as a chess problem (Chi, Glaser, and Rees, 1982). Fifth, the term “devices” contrasts the objects involved in IPS to natural objects, (social) events and so on, which are not the focus of IPS even though they can be formally modelled within the definition of IPS. It specifies for example, that IPS does not deal with problems requiring other abilities subsumed under the label emotional intelligence (Otto and Lantermann, 2005). Sixth, “by means of interacting” characterises the mode of exploration: information is generated and retrieved not by reading a manual or by applying content knowledge picked up in school, but by actively dealing with the system. Not explicitly included in the definition is the aspect of metacognition such as monitoring and reflecting one’s own problem-solving activities. But as problem solving is not only an ability but also a process, it needs some feedback and control structures that guide the activities. In the definition mentioned above it is implicitly assumed that these control processes accompany the entire problem-solving process (OECD, 2014; Wirth, 2004). IPS is embedded into cultural contexts, which is easily seen by the examples that are largely taken from industrialised countries. This cultural dependency is readily acknowledged even though it is assumed that IPS will be important in any society that wants to move to success in the sense of industrialisation and in the sense of a technological society (OECD, 2014).

Measuring interactive problem solving

With a definition of IPS at hand, the question of how to measure the construct immediately emerges. That is, how can the theoretical concept be adequately translated into empirical scales?

This question is far from trivial – besides the theoretical concept further psychometric and assessment demands need to be thoroughly met. According to Buchner (1995), two approaches are frequently used to measure IPS: 1) computer-simulated microworlds with a touch of real life composed of a large amount of variables (> 1 000 in the case of the famous scenario “Lohhausen” from Dörner, 1980); 2) simplistic, artificial and yet complex problems following certain construction rules (e.g. the DYNAMIS-approach using linear structural equation systems from Funke, 1992). Whereas the first approach uses ad hoc constructed scenarios to demonstrate individual differences, the second approach uses systematically constructed scenarios to demonstrate the effects of system attributes (Buchner, 1995). Both approaches have specific advantages and disadvantages in terms of 1) time used for testing; 2) realism; 3) the underlying measurement model; 4) available data on reliability and validity; 5) comparability between different scenarios; and 6) overall scalability with respect to varying difficulty levels. A detailed description of those is found in Greiff (2012).

Recently, Greiff et al. (2012) and Wüstenberg et al. (2012) proposed the MicroDYN approach as a measurement advancement of the DYNAMIS-approach by Funke (1992, 2001). MicroDYN was developed with a psychometric perspective for use in large-scale assessments (such as PISA) with computer-based test administration (Reeff and Martin, 2008). It contains an entire set of tasks, each of which consists of a small system of causal relations, which are to be explored within 3-4 minutes and afterwards controlled for given goal states. That is, MicroDYN allows for the assessment of the two overarching processes of knowledge acquisition and knowledge application. The main feature of the MicroDYN approach is the search for minimal complex systems (MICS), that is, systems which contain all (or at least most) of the features of a complex system (complexity, dynamics, polytely and intransparency; see Funke, 1991) but at the same time have low values for these parameters compared to the extremely difficult microworlds. From a psychometricians’ point of view this approach has several advantages over the microworlds used before, such as maintaining the validity of the tasks and reducing testing time to a minimum (compare also Greiff, Fischer, Stadler and Wüstenberg, 2015, for a detailed description of this approach, which they call the *multiple* complex system (MCS) approach, using the term multiple to stress the number of independent complex problem-solving tasks that are administered, rather than the level of complexity as in minimal complex systems; of note, the terms minimal and multiple complex systems are often used interchangeably).

The MicroDYN approach uses the formalism of linear structural equations (LSEs) to model systems with continuous variables (Funke, 1993). It is described in detail in Greiff et al. (2015) and in Wüstenberg et al. (2012). In this chapter, we argue not only for the use of minimal complex systems within the MicroDYN approach but also to extend the idea from systems with continuous variables to systems with discrete variables (MicroFIN; introduced by Buchner and Funke, 1993). Funke (2001) showed that both formal frameworks – LSEs and finite state automata (FSAs) – are ideal instruments for problem-solving research. Using minimal complex systems with the framework of finite state automata therefore seems a natural extension of the MicroDYN approach and has been introduced in greater detail by already Greiff et al. (2015). But before we go into the details of MicroDYN and MicroFIN, we will briefly explain the guiding philosophy behind the approach of minimal complex systems.

The philosophy behind minimal complex systems

One of the key concepts for the MicroDYN and MicroFIN approach is the overarching idea of "minimal complex systems" (MICS). The starting point for MICS is the assumption that complex systems are needed in problem-solving assessment because their features differ markedly from simple systems (in terms of complexity, connectivity, dynamics, intransparency and polytely; Dörner, 1986; Fischer et al., 2012) and are more than the sum of simple processes (Funke, 2010).

In the initial phase of problem-solving research beginning in the 1970s, it was thought that complex systems should realise a maximum amount of these features of complex problems, or, in other words, the more complex, the more connected, the more intransparent, and the more dynamic a system was, the better it was assumed to capture problem-solving behaviour. The underlying idea was to create computer simulations that were able to resemble reality to a highly detailed degree and thereby, to finally bring reality to the laboratory. So, for example, the famous microworld, Lohhausen (Dörner, 1980), contained over 1 000 variables with a testing time of well over several hours, whereas other less famous scenarios are said to incorporate up to 25 000 variables. The rationale behind this research strategy was to realise complexity in its extreme value, that is, to realise "maximal complex systems" (MACS).

What happened to these MACS systems from an assessment perspective? Despite their potential to realise highly complex systems they were associated with a lot of methodological problems and could not answer fundamental questions satisfactorily (Funke and Frensch, 2007): how to evaluate participants' actions and decisions (search for the optimal intervention); how to separate the effects of individual actions from those inherent in the dynamics of the microworld (one single intervention into a system with 5 000 connected variables can have 5 000 direct consequences and many more indirect effects in the next cycles); how to construct a test that contains more than one task because single tasks lead to dependencies between all actions within this system (Greiff et al., 2015); how to produce multiple tasks without spending too much time on assessment?

Here, basic measurement issues should have come into play early on, but conceptual considerations were always deemed more important than psychometric desiderata, which should form the basis of any test development.

The conception of MICS addresses these unsolved issues with a simple strategy: instead of realising increasingly complex systems (trying to aim for the top level of complexity) with a focus on content validity and psychometrically questionable results, MICS focuses on the lowest level and asks for the minimum value of complexity that is still valid in terms of representing the underlying concept of interactive problem solving. Complexity is a fuzzy term (Fischer et al., 2012; Funke, 2001) – we do not know what the most complex system on earth or even in the universe is.

So, the upper limit of complexity is still open. But, for good reasons, the lower limit of complexity must be somewhere between nothing and a little bit of complexity. Instead of searching for the upper bounds of complexity in MACS, in MICS we concentrate on the lower bound.

This shift in focus has been frequently applied in cognitive assessments; for example, intelligence tests are not meant to reflect real world tasks but to capture the basic processes necessary for mastering these tasks on an abstract level. It brings several advantages for test developers: 1) the time spent on a single scenario is not measured in hours but in minutes, thereby increasing the efficiency of test application; 2) due to the short time for task application, a series of independent tasks can be presented instead of one, thereby increasing reliability and making the set of tasks in principle scalable; 3) because of the existence of a formal framework for task development, tasks can be embedded in arbitrary real-world scenarios that are independent of the system's structure, thereby increasing ecological validity; and, 4) easy, medium and difficult tasks can be presented, broadening the range of difficulty and thereby increasing conceptual validity.

After more than 30 years of research with interactive problems, the measurement perspective now becomes a leading force for innovation. It continues a line of psychometric research started by Wagener (2001) and Kröner et al. (2005) who both tried to develop tests for interactive problem solving with the measurement perspective in mind. The MICS approach presented here is a logical continuation of this approach.

The basic elements of minimal complex systems

The MICS principle can be applied to the two formalisms mentioned above, linear structural equations and finite state automata (see again Greiff et al., 2015 for a detailed description). More specifically, the widely applied principles of LSEs and FSAs in the MACS approach are altered to allow a lower level of complexity by simply reducing the number of elements and connections involved and by reducing the time on each task. Because LSEs and FSAs (Funke, 2001) address different aspects (the first describes quantitative changes in continuous variables, the second qualitative changes in discrete variables), the following section introduces them separately even though it is assumed that they both tap into the same construct (OECD, 2014).

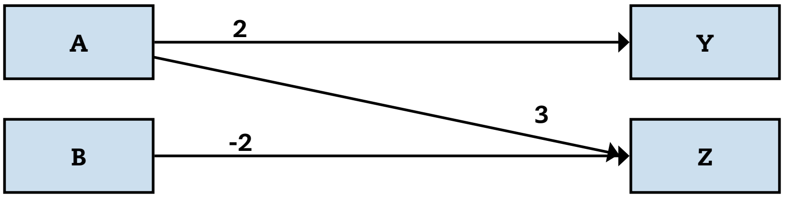
MicroDYN

In MicroDYN, a system's basic elements are the number of exogenous and endogenous variables and the number and type of relations between them. The relation between exogenous and endogenous variables can be qualified by its strength (weight of the relation) and direction (positive or negative). All of these features are used to scale a task's difficulty. In the MACS approach there may be 20 exogenous and endogenous variables with 30 connections between them, whereas the MICS approach gets by with as few as 4 to 6 variables and between 1 and 5 connections, but still yields systems with considerably varying difficulty. That is, even with this amount of formally low complexity, difficult tasks can easily be created (Greiff et al., 2012; Wüstenberg et al., 2012). The variables are labeled either abstractly with "A", "B", and "C" or with semantic meaningful labels such as "water", "wind", and "energy". Figure 6.1 illustrates a MICS system with two exogenous (A, B) and two endogenous (Y, Z) variables.

The subscript of each variable indicates the point in time (*t* respective *t+1*), the system itself being event driven in discrete steps. From Equation 1 it follows that the value of Y at time *t+1* is equal to the value of variable A at time *t*, multiplied by 2. From Equation 2 follows the same logic of computation. The graphical depiction in Figure 1 and the two equations (1) and (2) are equivalent in terms of the underlying causal structure but the diagram is more convenient to understand.

Exploration in the MicroDYN environment requires a careful analysis of the intervention effects: the increase of an exogenous variable could lead to an increase in one or more endogenous variables, or a decrease, or a mixed effect (increase in some variables and decrease in others), or to no effect at all. By carefully designing the sequence of interventions, a meaningful pattern of outputs can be generated and hypotheses about causal effects can be formulated (compare Sloman, 2005). As well as these relations between input variables (A and B in the example) and the output variables (Y and Z), the system may change by itself, which is expressed by connections between one endogenous variable and another (for example, Y could influence Z) or on itself (for example, Y could influence itself over time), and which also need to be represented. After the problem solver has spent some time gathering and expressing knowledge about the system, externally given goal values need to be reached. Thus, knowledge acquisition and knowledge application, the two main problem-solving processes (Novick and Bassok, 2005) can be assessed. Further developments incorporating the more detailed approach of Dörner (1986) with his five dimensions also exist and allow for a more detailed diagnostic approach (compare Fischer et al., 2012; Greiff, 2012).

Figure 6.1 **Structure of a MICS system with two exogenous and two endogenous variable2**



Note: The endogenous variables (A and B) and the exogenous variable are connected with causal paths indicated by arrows; weights for these paths are indicated next tot he arrows.

Source: Funke, J. (2001), "Dynamic systems as tools for analysing human judgement", Thinking and Reasoning, Vol. 7/1, pp. 69-89.

Formally, the system shown in Figure 6.1 can be described by two linear equations:

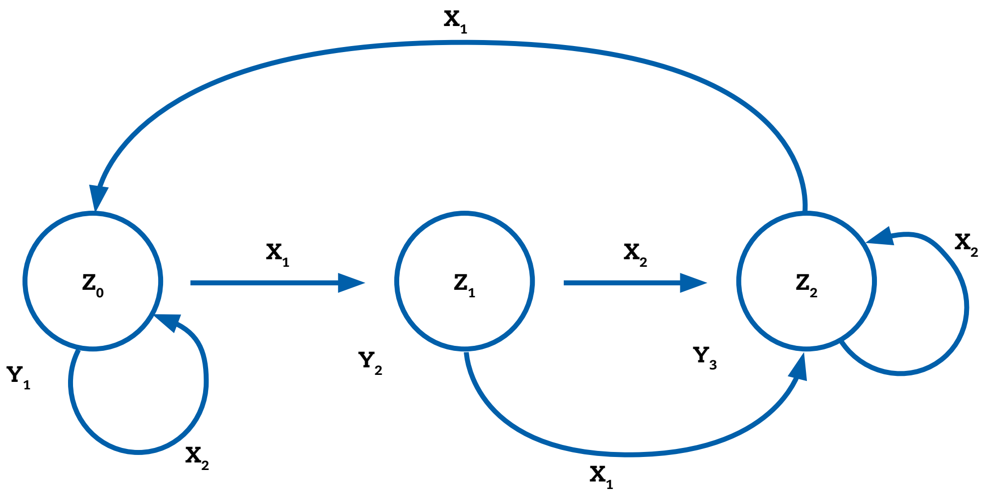
$$Y_{t+1} = 2 \times A_t \qquad (1)$$

$$Z_{t+1} = 3 \times A_t - 2 \times B_t \qquad (2)$$

MicroFIN

In MicroFIN, the basic elements are input signals, output signals, states and the state transitions between them. In contrast to MicroDYN, numbers and quantitative values are not important in MicroFIN, but elements of the problem are connected by qualitative transitions between them (Funke, 2001; Greiff et al., 2015). Figure 6.2 illustrates a simple finite state automaton with two input signals and three states (Thimbleby, 2007).

Figure 6.2 **A simple finite state automaton**



Note: The figure represents a finite state automaton with three states z_0, z_1, and z_2, and two input variables x_1 and x_2, leading to three possible output signals y_1, y_2, and y_3 depending on the reached state.

Source: Funke (2001:78).

Table 6.1 **State transition matrix of a fictitious finite state automaton**

state / output	**input**	
	$\mathbf{x_1}$	$\mathbf{x_2}$
z_0 / y_1	z_1	z_0
z_1 / y_2	z_2	z_2
z_2 / y_3	z_0	z_2

Note: The table shows the state transition matrix for the finite state automaton represented in Figure 6.2. For each state there exists exactly one corresponding output signal (y_1, y_2, and y_3). The cells of the matrix indicate the next state for the machine given the combination of current state and input signal.

Source: Funke (2001:77).

Formally, the diagram in Figure 6.2 is equivalent to the state transition matrix in Table 6.1. Once again, the graphical representation seems more convenient when representing an FSA. Acquiring knowledge in a MicroFIN environment requires a step-by-step analysis of the state transitions. In principle, at least as many steps are required as there are different state transitions before an FSA is entirely penetrated, whereas in practice, this number can be reduced by detecting hierarchies and analog functions within the automaton. Wirth (2004) suggests first ideas for measuring exploration behaviour, as well as knowledge representation and its application, whereas Neubert, Kretzschmar, Wüstenberg and Greiff (2015) present first empirical results on the validity of MicroFIN as an empirically applicable assessment instrument.

To give an example of an FSA, a digital watch is (in large part) a typical finite state machine. Transitions between different states are mostly of a qualitative nature (e.g. from alarm mode to snooze mode) and the tasks mentioned above and also present in MicroDYN have to be met: knowledge about the system has to be gathered and represented (knowledge acquisition) and a given goal, such as setting the alarm for 7 o'clock in the morning, has to be reached (knowledge application). Whereas MicroDYN tasks yield the advantage of being homogenous and adding up to a narrow but reliable test set (low bandwidth, high fidelity), MicroFIN tasks are closer to real life and more heterogeneous in the skills they test (high bandwidth, low fidelity; Greiff et al., 2015). Besides these differences, there are considerable overlaps between MicroDYN and MicroFIN, which we will elaborate on now.

Common elements of MicroDYN and MicroFIN

MicroDYN and MicroFIN have some common elements with respect to the task subjects have to meet and with regard to the potential diagnostic procedures. From the subjects' point of view, both paradigms require similar activities, at least in principle: the identification of an unknown dynamic device and the subsequent control of this device. Subjects have to set up an exploration strategy which generates information about the system and incorporate this into a mental representation (knowledge acquisition); the subsequent control performance is the result of a goal-directed application of the acquired knowledge (knowledge application). Within these two main processes (Novick and Bassok, 2005), further subprocesses can be distinguished and separate indicators for the problem-exploration stage have yielded promising results (see below). Because of the similarity of the tasks in MicroDYN and MicroFIN, we derive the same diagnostic parameters in both cases. A detailed description of potential indicators and scoring procedures is found in Buchner (1995), Neubert et al. (2015), and Wüstenberg et al. (2012). In general, both tasks tap into the same construct, as empirically shown by Greiff et al. (2013) and Neubert et al. (2015). For the first time, MicroDYN and MicroFIN make psychometric assessment procedures available to capture this construct originally coming from cognitive psychology.

Recent results on interactive problem solving

Empirical research on MICS employing the MicroDYN and MicroFIN approach has gathered momentum over the last couple of years with a number of interesting findings now available. Most of these studies have demonstrated the gain in validity from combining formal frameworks from problem-solving research with a psychometric measurement perspective. Knowledge acquisition and knowledge application as overarching processes can be empirically separated in different populations (e.g. Greiff et al., 2013; Kröner et al., 2005; Wüstenberg, et al., 2012). The correlation between the different problem-solving processes is usually high, but different from unity. Even more important than concurrence between empirical and theoretical structure is that MicroDYN and MicroFIN as well as other measures of problem solving are correlated with and yet distinct from general intelligence (Kröner et al. 2005; Wüstenberg et al., 2012; Sonnleitner et al., 2012) and from working memory (Bühner, Kröner and Ziegler, 2008; Schweizer et al., 2013). Further, IPS predicts school achievement incrementally beyond intelligence, but only if MICS are used rather than MACS, showing that 1) measures of intelligence do not fully overlap with IPS; and that 2) the MICS approach is needed to produce reliable problem-solving scales (compare in detail Greiff et al., 2012, comparing MicroDYN to Space Shuttle; Greiff, Stadler, Sonnleitner, Wolff and Martin, 2015 comparing MicroDYN and MicroFIN to the Tailorshop, another assessment instrument based on MACS). Abele et al. (2012) report that IPS is predictive of building up knowledge in specific domains, which leads to domain-bound problem-solving skills. That is, domain-specific problem solving is largely determined by knowledge in the respective area, but IPS might be a significant prerequisite for gaining this knowledge. In general, MicroDYN and MicroFIN exhibit good psychometric characteristics and promising results with regard to internal and external validity (for further studies see, for instance, Fischer et al., 2015; Frischkorn et al., 2014; Greiff and Wüstenberg, 2014; Wüstenberg, Greiff, Molnár and Funke, 2014). For a more comprehensive example of using MICS for empirical research, see Chapter 8.

Discussion

MicroDYN and MicroFIN focus on the use of minimal complex systems for the assessment of IPS. The existing body of empirical research supports this approach. This final section addresses three open questions. First, can IPS be separated conceptually and empirically from constructs such as intelligence, knowledge or working memory capacity? Second, are knowledge acquisition and knowledge application separate and yet central features of problem solving? And third, have minimal complex systems the same potential for assessing the construct of dealing with complexity as maximal complex systems?

Construct validity

Obviously, problem solving and intelligence are overlapping constructs and there are at least two positions claiming that they are considerably related. First, the ability to solve problems features prominently in almost any definition of human intelligence. Thus, problem-solving capacity is viewed as one component of intelligence (Funke and Frensch, 2007). However, for Dörner (1986), it is the other way around: he considers intelligence as operationalised in classical assessment instruments to be one of several components of problem solving, proposing a theory of operative intelligence combining both constructs on a conceptual level. Second, intelligence is often assumed to be an empirical predictor of problem-solving ability. In a recent meta-analysis, Stadler et al. (2015) corroborate this finding and report that across 47 studies the relation between problem solving and intelligence was moderate (with a mean correlation of 0.43). In summary, the available evidence suggests that the concepts of intelligence and problem solving are moderately related, whereas specific subcomponents of intelligence and problem solving might share additional variance. However, when predicting external criteria, problem solving explains some unique variance in the criteria. The existing empirical evidence does not speak, however, to the issue of whether subcomponents of intelligence predict subcomponents of problem solving or whether the opposite causal relation holds. Overall, there is still more research needed on the connection between intelligence and problem solving (Wittmann and Süß, 1999) and we are confident that researchers will turn to this issue even more intensively in the future.

Knowledge acquisition and knowledge application

Why should we concentrate on these particular two dimensions as central elements of problem solving? The PISA 2012 framework on problem solving, arguably the largest student assessment worldwide, differentiates four groups of problem-solving processes (OECD, 2014): 1) exploring and understanding; 2) representing and formulating; 3) planning and executing; and 4) monitoring and reflecting. The first three we understand to be mostly equivalent to knowledge acquisition (which includes exploration of the problem situation), and knowledge application, with the main difference that our proposed terms are closer to overt behavior than the PISA terms. The fourth is a meta-component, which is difficult to assess and largely targeted by questionnaires (OECD, 2014; Wirth, 2004).

Another argument why the two dimensions of knowledge acquisition and knowledge application should be seen as central features of problem solving comes from studies with primates. The *Werkzeuggebrauch* (use of tools) of chimpanzees consists in exploring a certain device for later use in a problem situation. Even for humans, *intelligenter Werkzeuggebrauch* (intelligent use of tools) seems to be one of the oldest approaches to problem solving, which Jonassen (2003) now calls the use of cognitive tools (exploration and knowledge acquisition). Whereas nearly 100 years ago the chimpanzees of Wolfgang Köhler (1921) on the island of Tenerife used wooden sticks for reaching out to bananas, we nowadays use (electronic) devices to reach different goals in nearly every area of our life adhering to the application of knowledge. Thus, it may well be that the two main processes, knowledge acquisition and knowledge application, may be augmented by additional processes such as, a successful exploration of the problem as a prerequisite to acquire knowledge. Research will show, which of the competing theoretical conceptions draw the closest picture of the empirical reality (Fischer et al., 2012).

Minimal versus maximal complex systems

Certainly, a microworld such as Lohhausen (Dörner, 1980), with over 1 000 variables, provides a different experience for participants than a small-scale MicroDYN or MicroFIN scenario. The overwhelming amount of information in the case of Lohhausen, the high-stakes situation for the problem solver, the extremely knowledge-rich context and its associated manifold strategies make a substantial difference. For future research on long-term strategy development, large-scale scenarios will have their justification. On the other hand, the use of MicroDYN and MicroFIN tasks offer huge advantages for assessment: reliable measurement, varying difficulty and psychometric scalability, to name just a few. At the same time, the difficulty of the tasks as well as their semantic content can be easily adapted to different target populations, which helps to increase content validity. This flexibility, even though it is accompanied with a narrowing down of the processes and the conceptualisation targeted in the assessment, should encourage us to explore the potential gains of minimal complex systems more deeply.

No doubt, using MICS instead of MACS comes at a cost: the cost of narrowing interactive problem solving down to minimal problems and thereby somewhat impairing the measures' external validity. MACS assessment devices on the other hand are associated with an even larger impairment: the neglect of fundamental assessment principles (like a reference to a "best" solution). That is, when assessing interactive problem solving, a viable compromise needs to be found and considering theoretical and empirical evidence available today, the compromise offered by MICS is one worth while pursuing.

Notes

[1] Please note that in the research literature the terms complex, dynamic, interactive and sometimes creative problems are used rather inconsistently and interchangeably. For the sake of clarity, we will consistently use the term interactive problem solving.

References

Abele, S. et al. (2012), "Dynamische Problemlösekompetenz: Ein bedeutsamer Prädiktor von Problemlöseleistungen in technischen Anforderungskontexten" [The importance of general cognitive determinants in mastering job demands. Some research on the example of dynamic and technical problem solving], *Zeitschrift für Erziehungswissenschaft*, Vol. 15/2, pp. 363-391.

Buchner, A. (1995), "Basic topics and approaches to the study of complex problem solving", in P.A. Frensch and J. Funke (eds.), *Complex Problem Solving: The European Perspective,* Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 27-63.

Buchner, A. and J. Funke (1993), "Finite-state automata: Dynamic task environments in problem-solving research", *Quarterly Journal of Experimental Psychology*, Vol. 46A/1, pp. 83-118.

Bühner, M., S. Kröner and M. Ziegler (2008), "Working memory, visual–spatial intelligence and their relationship to problem solving", *Intelligence*, Vol. 36/6, pp. 672-680.

Chi, M.T.H., R. Glaser and E. Rees (1982), "Expertise in problem solving", in R.J. Sternberg (ed.), *Advances in the Psychology of Human Intelligence, Volume 1*, Erlbaum, Hillsdale, NJ, pp. 7-75.

Dörner, D. (1996), *The Logic of Failure: Recognizing and Avoiding Error in Complex Situations*, Addison-Wesley, Reading, MA.

Dörner, D. (1986), "Diagnostik der operativen Intelligenz" [Assessment of operative intelligence], *Diagnostica*, Vol. 32/4, pp. 290-308.

Dörner, D. (1980), "On the difficulties people have in dealing with complexity", *Simulation & Gaming*, Vol. 11, pp. 87-106.

Fischer, A., S. Greiff and J. Funke (2012), "The process of solving complex problems", *Journal of Problem Solving*, Vol. 4/1, pp. 19-42, http://dx.doi.org/10.7771/1932-6246.1118.

Fischer, A., S. Greiff, S. Wüstenberg, J. Fleischer, F. Buchwald and J. Funke (2015), "Assessing analytic and interactive aspects of problem solving competency", *Learning and Individual Differences*, Vol. 39, pp. 172-179, http://dx.doi.org/10.1016/j.lindif.2015.02.008.

Frischkorn, G., S. Greiff and S. Wüstenberg (2014), "The development of complex problem solving: A latent growth curve analysis", *Journal of Educational Psychology*, 106/4, pp. 1007-1020.

Funke, J. (2010), "Complex problem solving: A case for complex cognition?", *Cognitive Processing*, Vol. 11/2, pp. 133-142.

Funke, J. (2001), "Dynamic systems as tools for analysing human judgement", *Thinking and Reasoning*, Vol. 7/1, pp. 69-89.

Funke, J. (1993), "Microworlds based on linear equation systems: A new approach to complex problem solving and experimental results", in G. Strube and K.-F. Wender (eds.), *The Cognitive Psychology of Knowledge*, Elsevier Science Publishers, Amsterdm, pp. 313-330.

Funke, J. (1992), "Dealing with dynamic systems: Research strategy, diagnostic approach and experimental results", *German Journal of Psychology*, Vol. 16/1, pp. 24-43.

Funke, J. (1991), "Solving complex problems: Exploration and control of complex systems", in R.J. Sternberg and P.A. Frensch (eds.), *Complex Problem Solving: Principles and Mechanisms,* Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 185-222.

Funke, J. and P.A. Frensch (2007), "Complex problem solving: The European perspective – 10 years after", in D.H. Jonassen (ed.), *Learning to Solve Complex Scientific Problems*, Lawrence Erlbaum, New York, pp. 25-47.

Greiff, S. (2012), *Individualdiagnostik komplexer Problemlösefähigkeit* [Assessment of Complex Problem-Solving Ability], Waxmann, Munster.

Greiff, S., et al. (2015), Assessing complex problem solving skills with Multiple Complex Systems., *Thinking & Reasoning*,Vol. 21, pp. 356-382.

Greiff, S., et al. (2013), "A multitrait-multimethod study of assessment instruments for complex problem solving", *Intelligence*, Vol. 41/5, pp. 579-596.

Greiff, S. and J. Funke (2009), "Measuring complex problem solving: The MicroDYN approach", in F. Scheuermann and J. Björnsson (eds.), *The Transition to Computer-Based Assessment: New Approaches to Skills Assessment and Implications for Large-Scale Testing,* Office for Official Publications of the European Communities, Luxembourg, pp. 157-163.

Greiff, S., D.V. Holt and J. Funke (2013), "Perspectives on problem solving in cognitive research and educational assessment: Analytical, interactive, and collaborative problem solving", *Journal of Problem Solving*, Vol. 5, pp. 71-91.

Greiff, S., M. Stadler, P. Sonnleitner, C. Wolff and R. Martin (2015), "Sometimes less is more: Comparing the validity of complex problem solving measures", *Intelligence*, Vol. 50, pp. 100-113.

Greiff, S. and S. Wüstenberg (2014), "Assessment with microworlds: Factor structure, invariance, and latent mean comparison of the MicroDYN test", *European Journal of Psychological Assessment*, Vol. 30/1, pp. 1-11.

Greiff, S., S. Wüstenberg and J. Funke (2012), "Dynamic problem solving: A new assessment perspective", *Applied Psychological Measurement*, Vol. 36/3, pp. 189-213. http://dx.doi.org/10.1177/0146621612439620.

Greiff, S., et al. (2013), "Complex problem solving in educational settings – Something beyond *g*: Concept, assessment, measurement invariance, and construct validity, *Journal of Educational Psychology*, Vol. 105/2, pp. 364-379, http://dx.doi.org/10.1037/a0031856.

Jonassen, D.H. (2003), "Using cognitive tools to represent problems", *Journal of Research on Technology in Education*, Vol. 35/3, pp. 362-381.

Köhler, W. (1921), *Intelligenzprüfungen an Menschenaffen* [Examining intelligence in primates], Springer, Berlin.

Kröner, S., J.L. Plass and D. Leutner (2005), "Intelligence assessment with computer simulations", *Intelligence*, Vol. 33/4, pp. 347-368, http://dx.doi.org/10.1016/j.intell.2005.03.002.

Mayer, R.E. and M.C. Wittrock (2006), "Problem solving", in P.A. Alexander and P.H. Winne (eds.), *Handbook of Educational Psychology*, Lawrence Erlbaum, Hillsdale, NJ, pp. 287-303.

Neubert, J.C., et al. (2015), "Extending the assessment of complex problem solving to finite state automata: Embracing heterogeneity", *European Journal of Psychological Assessment*, Vol. 31, pp. 181-194.

Novick, L.R. and M. Bassok (2005), "Problem solving", in K.J. Holyoak and R.G. Morrison (eds.), *The Cambridge Handbook of Thinking and Reasoning*, Cambridge University Press, pp. 321-349.

OECD (2014), *PISA 2012 Results: Creative Problem Solving (Volume V): Students' Skills in Tackling Real-Life Problems*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264208070-en.

Otto, J.H. and E.-D. Lantermann (2005), "Individual differences in emotional clarity and complex problem solving", *Imagination, Cognition and Personality*, Vol. 25/1, pp. 3-24.

Reeff, J.-P. and R. Martin (2008), "Use of the internet for the assessment of students' achievement", in J. Hartig, E. Klieme and D. Leutner (eds.), *Assessment of Competencies in Educational Settings*, Hogrefe & Huber, Göttingen.

Schweizer, F., S. Wüstenberg and S. Greiff (2013), "Validity of the MicroDYN approach: Complex problem solving predicts school grades beyond working memory capacity", *Learning and Individual Differences*, Vol. 24, pp. 42-52.

Sloman, S. (2005), *Causal Models. How People Think About the World and its Alternatives*, Oxford University Press.

Sonnleitner, P. et al. (2012), "The *Genetics Lab*: Acceptance and psychometric characteristics of a computer-based microworld assessing complex problem solving", *Psychological Test and Assessment Modeling*, Vol. 54/1, pp. 54-72.

Stadler, M., N. Becker, M. Gödker, D. Leutner and S. Greiff (2015), "Complex problem solving and intelligence: A meta-analysis", *Intelligence*, Vol. 53, pp. 92-101.

Thimbleby, H. (2007), *Press On: Principles of Interaction Programming*, MIT Press, Cambridge, MA.

Wagener, D. (2001), *Psychologische Diagnostik mit komplexen Szenarios: Taxonomie, Entwicklung, Evaluation* [Psychological Assessment with Complex Scenarios:Taxonomy, Development, Evaluation], Pabst Science Publishers, Lengerich.

Wirth, J. (2004), *Selbstregulation von Lernprozessen* [Self-Regulation of Learning Processes], Waxmann, Münster.

Wittmann, W.W. and H.-M. Süß (1999), "Investigating the paths between working memory, intelligence, knowledge, and complex problem solving performances via Brunswik symmetry", in P.L. Ackerman, P.C. Kyllonen and R.D. Roberts (eds.), *Learning and Individual Differences: Process, Trait, and Content Determinants*, American Psychological Association, Washington, DC, pp. 77-108.

Wüstenberg, S., S. Greiff and J. Funke (2012), "Complex problem solving: More than reasoning?", *Intelligence*, Vol. 40/1, pp. 1-14, http://dx.doi.org/10.1016/j.intell.2011.11.003.

Wüstenberg, S., et al. (2014), "Cross-national gender differences in complex problem solving and their determinants", *Learning and Individual Differences*, Vol. 29, pp. 18-29, http://dx.doi.org/10.1016/j.lindif.2013.10.006.

Chapter 7

The history of complex problem solving

By

Andreas Fischer

Forschungsinstitut Betriebliche Bildung (f-bb), Nuremberg, Germany.

Samuel Greiff

Maison des Sciences Humaines, Université du Luxembourg, Luxembourg.

and Joachim Funke

Department of Psychology, Heidelberg University, Germany.

Complex problem solving (CPS) is about reaching one's goals taking into account a large number of highly interrelated aspects. CPS has a rich history in experimental and psychometric research. The chapter highlights some of the most important findings of this research and shows its relationship to interactive problem solving. More specifically, it 1) characterises typical human strategies and shortcomings in coping with complex problems; 2) summarises some of the most influential theories on cognitive aspects of CPS; and 3) outlines the history of including CPS skills and competency in assessment contexts. The last section summarises the current state of play and points out some trends for future research.

Correspondence concerning this chapter should be addressed to Joachim Funke, Department of Psychology, Heidelberg University, Hauptstraße 47–51, D–69117 Heidelberg, Germany. This research was supported by a grant of the Fonds National de la Recherche Luxembourg (ATTRACT "ASSKI21") to the second author and by grants of BMBF (FKZ: 01JG1062) and DFG (Az. Fu 173/14) to the last author.

Introduction

Research on complex problem solving (CPS) arose in the 1970 (e.g. Dörner, 1975; Dörner, Drewes and Reither, 1975) when real problems like the limits of growth (Meadows, Meadows, Zahn and Milling, 1972) and the oil crisis of 1973 made it increasingly clear that we needed more than traditional reasoning tasks to examine the ways humans solve problems and make decisions in naturalistic complex and dynamic situations (Brehmer, 1992; Klein, Orasanu, Calderwood and Zsambok, 1993; Gonzalez, Vanyukov and Martin, 2005). Recent examples of complex problems are found in decision making about global climate politics (e.g. Amelung and Funke, 2013), in avoiding bankruptcy for mismanaged corporations (e.g. Funke, 2003), or in coping with nuclear disasters like Fukushima and their consequences (e.g. Dörner, 1997).

Formally speaking CPS can be defined "as (a) knowledge acquisition and (b) knowledge application concerning the goal-oriented control of systems that contain many highly interrelated elements (i.e., complex systems)" (Fischer, Greiff and Funke, 2012:19). This definition is very broad and includes "interactive problem solving" as one aspect of controlling complex and dynamic systems. More than 20 years ago, Frensch and Funke defined CPS as follows:

> "CPS occurs to overcome barriers between a given state and a desired goal state by means of behavioral and/or cognitive, multi-step activities. The given state, goal state, and barriers between given state and goal state are complex, change dynamically during problem solving, and are intransparent. The exact properties of the given state, goal state, and barriers are unknown to the solver at the outset. CPS implies the efficient interaction between a solver and the situational requirements of the task, and involves a solver's cognitive, emotional, personal, and social abilities and knowledge." (1995:18)

Since the early 1980s, the pioneers of CPS research have developed computer simulations of real complex[1] problems in order to examine learning and decision making under realistic circumstances in psychological laboratories (e.g. Berry and Broadbent, 1984; Dörner, Kreuzig, Reither and Stäudel, 1983). Some of the most famous computer simulations of complex problems were developed during this period, involving ruling a city (Lohhausen; Dörner et al., 1983), organising developmental aid (Moroland; Strohschneider and Güß, 1999), managing a tailorshop (Putz-Osterloh, 1981), controlling a sugar factory (Berry and Broadbent, 1987) or finding out about a complex toy (Big Trak; Klahr and Dunbar, 1988). In the years that followed, the simulation of complex problems was increasingly established as new research paradigm and the scientific community documented characteristic human strategies and shortcomings in coping with complex problems (e.g. Dörner, 1982). In the late 1980s, a set of theoretical contributions significantly expanded the information processing theory of human problem solving (Newell and Simon, 1972) by elaborating on aspects such as the role of hypothesis testing (Klahr and Dunbar, 1988) and implicit learning (Berry and Broadbent, 1987) in the process of solving complex problems (see Frensch and Funke, 1995; for a recent overview, see Fischer et al., 2012). Dörner (1986) proposed using CPS simulations for assessments, and initiated a lively debate on the incremental value of those simulations beyond traditional measures of intelligence (Funke, 2003). In the 1990s CPS simulations were increasingly criticised from a psychometric point of view: the different problems seemed to have little in common (Funke, 1992, 2001, 2010) and often conclusions about a person's ability were drawn from interventions in a single problem only (Greiff, 2012; Wittmann and Süß, 1999). The research community increasingly elaborated on how to compare different simulations and the effects of a system's features on a problem's difficulty, which was an important prerequisite for the psychometric approach to CPS (Funke, 2001). In the early 21st century, the psychometric quality of CPS simulations was rehabilitated (Danner et al., 2011a; Wüstenberg, Greiff and Funke, 2012) and CPS skills and performance proved to be an important complement to traditional measures of intelligence in large-scale assessments such as the Programme for International Student Assessment (PISA) 2000, 2003 and 2012 (OECD, 2010; Fischer, 2015).

In the following sections we will outline some of the most important findings in the history of CPS research. This chapter will 1) characterise some typical human strategies and shortcomings in coping with complex problems; 2) summarise some of the most influential theories on important aspects of CPS; and 3) outline the history of CPS in assessment contexts. The final section summarises the current position and points to some trends for future research.

Human failures and strategies

In the early 1980s, the research group around Dörner, Kreuzig, Reither and Stäudel (1983) elaborated on some of the most frequent errors problem solvers make in complex situations, such as ignoring trends, underestimating exponential growth and thinking in causal chains instead of causal networks (Dörner, 1982). By comparing people who do well in CPS situations with those who do not, Dörner (1989, 1996) was able to determine some important aspects of problem-solving competency: On an aggregate level, good problem solvers made more decisions simultaneously than bad ones (and increasingly so in the course of problem solving), that is, they saw more possibilities to influence their situation. They also made more decisions per goal and took different aspects of the system into account. Good problem solvers focused on central causes early on, and did not gradually shift from trivial effects to important issues as bad problem solvers did.

Interestingly, good and bad problem solvers did not differ in the number of hypotheses they generated, but in the number they tested: good problem solvers tested their hypotheses and asked more questions about causal links behind events. Good problem solvers also were less often distracted by urgent but unimportant subproblems (contrast with “ad-hocism”; Dörner, 1996:25) and their decisions were more consistent over time, whereas bad problem solvers tended to change their current subproblem more often and were more easily distracted. Good problem solvers structured, reflected, criticised and modified their own hypotheses and behaviour to a greater degree, whereas bad problem solvers tended to delegate subproblems instead of facing them responsibly. Jansson (1994) reported seven erroneous responses to complex problem situations: 1) acting directly on current feedback; 2) insufficient systematisation; 3) insufficient control of hypotheses and strategies; 4) no self-reflection; 5) selective information gathering; 6) selective decision making; and 7) thematic vagabonding. For a more recent summary of human failures in complex situations, see Schaub (2006).

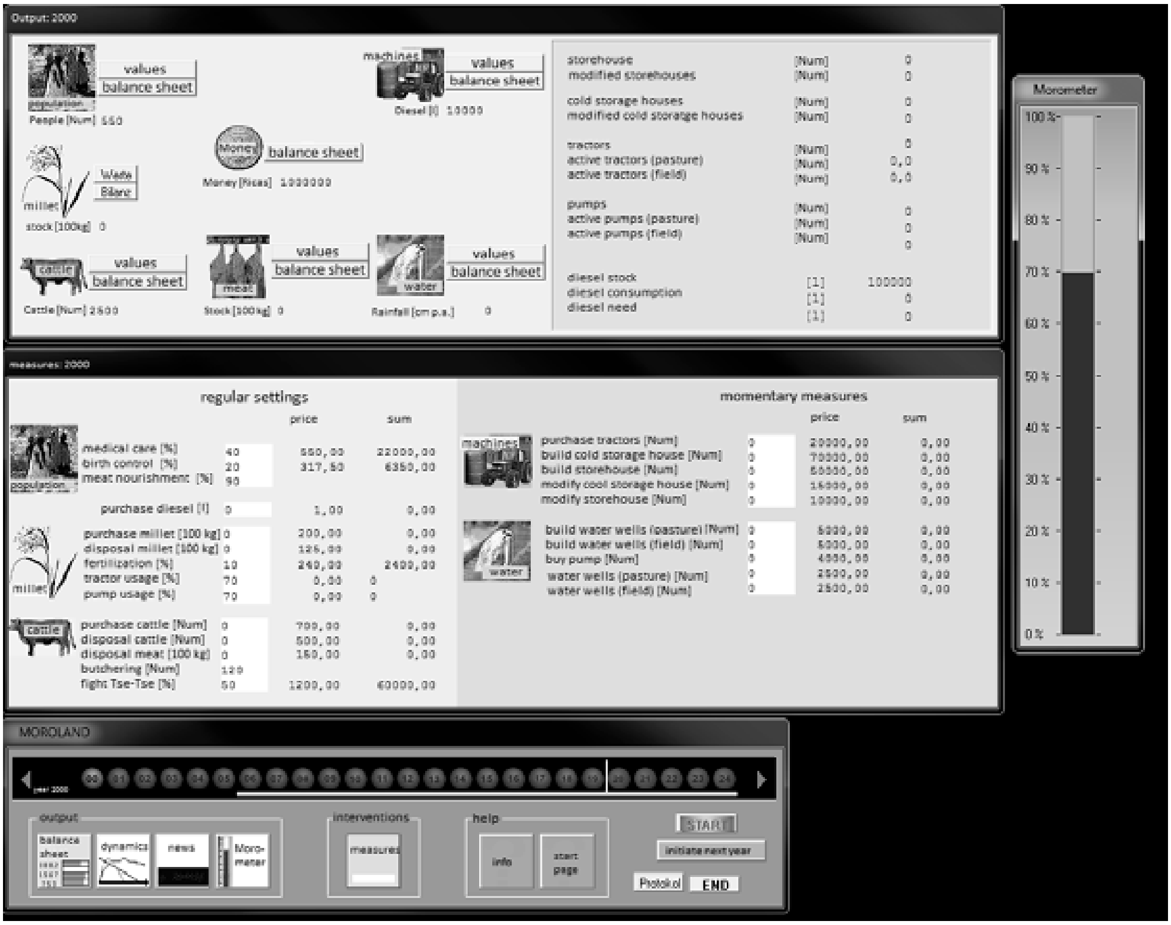
In earlier research, the behaviour of bad problem solvers was often explained by a lack of general intelligence, or insufficient domain-specific prior knowledge of subjects (compare Fischer et al., 2012), but according to Strohschneider and Güß (1999), recent evidence indicates a strong relationship with domain-general strategic knowledge (see below).

Based on the complex Moro problem (see Figure 7.1 for a modern implementation of the simulation), where the problem solver has to manage developmental aid for a small tribe of semi-nomads, Strohschneider and Güß (1999) defined a set of deficient strategic, operational and tactical patterns, which may indicate a lack of strategic knowledge:

- **incomplete exploration of domains**: the problem solver selectively considers certain problem areas and misses other aspects of the complex problem situation
- **feedback strategy**: the problem solver selectively reacts to signals emanating from the system – such as alarm messages or complaints
- **insufficient adaption of interventions**: the problem solver does not adapt his or her decisions – such as the annual sale of millet – to changes in the situation such as insufficient cattle
- **decisions without information**: the problem solver does not gather information relevant to his or her decisions such as the prices of goods

- **lack of effect control**: the problem solver fails to monitor the effects of decisions, for example not viewing population figures, birth rates or death rates after implementing medical services
- **collisions**: the outcomes of the problem solver's decisions cancel each other out.

Figure 7.1 **Screenshot of a simulation of the complex Moro problem**



Output: 2000

People [Num] 550

Diesel [l] 10000

Money [Ficas] 1000000

stock [100kg] 0

Cattle [Num] 2600

Stock [100 kg] 0

Rainfall [cm p.a.] 0

storehouse	[Num]	0
modified storehouses	[Num]	0
cold storage houses	[Num]	0
modified cold storatge houses	[Num]	0
tractors	[Num]	0
active tractors (pasture)	[Num]	0,0
active tractors (field)	[Num]	0,0
pumps	[Num]	0
active pumps (pasture)	[Num]	0
active pumps (field)	[Num]	0
diesel stock	[1]	100000
diesel consumption	[1]	0
diesel need	[1]	0

measures: 2000

regular settings

		price	sum
medical care [%]	40	550,00	22000,00
birth control [%]	20	317,50	6350,00
meat nourishment [%]	90		
purchase diesel [l]	0	1,00	0,00
purchase millet [100 kg]	0	200,00	0,00
disposal millet [100 kg]	0	125,00	0,00
fertilization [%]	10	240,00	2400,00
tractor usage [%]	70	0,00	0
pump usage [%]	70	0,00	0
purchase cattle [Num]	0	700,00	0,00
disposal cattle [Num]	0	500,00	0,00
disposal meat [100 kg]	0	150,00	0,00
butchering [Num]	120		
fight Tse-Tse [%]	50	1200,00	60000,00

momentary measures

		price	sum
purchase tractors [Num]	0	20000,00	0,00
build cold storage house [Num]	0	70000,00	0,00
build storehouse [Num]	0	50000,00	0,00
modify cool storage house [Num]	0	15000,00	0,00
modify storehouse [Num]	0	10000,00	0,00
build water wells (pasture) [Num]	0	5000,00	0,00
build water wells (field) [Num]	0	5000,00	0,00
buy pump [Num]	0	4000,00	0,00
water wells (pasture) [Num]	0	2500,00	0,00
water wells (field) [Num]	0	2500,00	0,00

Source: Lantermann et al. (2000), SYRENE: Umwelt- und Systemlernen mit Multimedia.

Building on these and other strategic shortcomings in coping with complex problems, many promising training approaches have been developed to foster problem-solving performance. For example, Dörner (1996) and Vester (2007) proposed computer simulations to learn CPS (compare with Kretzschmar and Süß, 2015). Hestenes (1992) proposed modelling instructions to teach a systematic way of building viable models of complex situations (continuously adapting them to empirical feedback) and D'Zurilla and Goldfried (1971) developed training for problem solving that proved to be helpful in a variety of contexts (Liebeck, 2008). Kluge (2008) provides a critical review of such training.

Cognitive theories on the process of solving complex problems

This section presents some of the most significant theories about important cognitive aspects[2] of CPS, which began to emerge in the late 1980s: a brief outline the theory of Klahr and Dunbar (1988) on scientific knowledge acquisition as well as an important expansion of Schunn and Klahr (2002) – an expansion that introduces the aspect of information reduction which is of special importance to CPS (Fischer et al., 2012). We will then look at theories of system control and knowledge application, namely Broadbent, Fitzgerald and Broadbent's theory (1986) about the role of explicit and implicit knowledge in CPS, and a recent expansion of Osman (2010) elaborating on the role of monitoring in CPS.

Theories on knowledge acquisition

Based on the ideas of Simon and Lea (1974), Klahr and Dunbar (1988) expanded the theory of human problem solving (Newell and Simon, 1972) – which described problem solving simply as the search for as solution (structured by methods such as means-end analysis, planning, hill-climbing, analogy or domain-specific knowledge) – in order to explain scientific discovery learning in the process of CPS. They described the endeavour of scientific discovery as a dual search. On the one hand, the problem solver has to search for experiments (i.e. empirical data) in order to generate and test hypotheses. On the other hand, the problem solver has to search for hypotheses (i.e. chunks of knowledge) that sufficiently explain the data gathered. With their dual-search theory, Klahr and Dunbar (1988) elaborated on an important aspect of the process of CPS (Funke, 2001). Later on, the theory was expanded by Schunn and Klahr (2002), who identified an additional search space, namely the search for data representations that highlight the elements that are crucial for finding a solution; they realised that in CPS people did not just search for hypotheses and experiments, but also for which set of variables to consider as relevant "data" at all. For example, problem solvers had to find out that the colour, form or size of houses didn't matter when explaining the route of a simulated milk truck. Thereafter, participants increasingly focused their knowledge acquisition on the relevant aspects of the problem and reduced irrelevant information (compare with the information reduction of Gaschler and Frensch, 2007; and the *Schwerpunktbildung* of Dörner et al., 1983). This is an important prerequisite for efficient planning in complex problem solving (Klauer, 1993; Fischer et al., 2012).

Theories on knowledge application

Broadbent, Fitzgerald and Broadbent (1986) researched the impact of explicit and implicit knowledge on performance in dynamic decision making: in their instance-theory of complex system control they pointed out that performance in a problem-solving situation can change without changes in explicit knowledge due to instances of the system (consisting of the current output and a corresponding input) being remembered instantly as a result of implicit learning. The theory poses that problem solvers get better at controlling dynamic systems by storing their reactions to the states perceived as an instance in a kind of "look-up-table" and by applying instances to similar situations (compare with Gonzalez, Lerch and Lebiere, 2003; Dienes and Fahey, 1995). But in addition to knowledge about the states of a system, there also can be knowledge about the system's structure, i.e. about the abstract rules governing the system's behaviour (Berry and Broadbent, 1987; Schoppek, 2002). With this latter kind of knowledge in mind, different heuristics, such as planning (e.g. Klauer, 1993), can be applied in the process of CPS (Fischer et al., 2012). Schoppek (2002) emphasised that in more complex systems the application of explicit structural knowledge may be of crucial importance for system control, because a tremendous amount of instances would be needed to control the system based on instances. The question of when and if a problem solver searches for either rules or instances to apply was elaborated on by the dual-search models of problem solving outlined above. For instance, the search for rule knowledge may increase with factors like the relevance of structure (Dienes and Fahey, 1995) or a lack of goal specificity (Vollmeyer, Burns and Holyoak, 1996).

But even when problem solvers know about both instances and rules, and apply them to the best of their knowledge, the effectiveness of an intervention (derived from explicit and/or implicit knowledge) cannot be taken for granted in a CPS situation. Particularly where prior knowledge is mostly false or incomplete (Dörner, 1997), the effectiveness of interventions has to be monitored. In her monitoring-control framework, Osman specified how individuals intensify monitoring as a result of experiencing uncertainty: "High uncertainty will lead to continuous monitoring of the task and goal directed actions, and low uncertainty will lead to periodic monitoring of the task and goal directed actions" (2010:73). Behaviour intended to control the environment generates feedback on one's hypotheses and plans, while monitoring the consequences of one's actions (self-monitoring) or the problem's dynamics (task-monitoring) allows for adjusting one's plans and hypotheses. Thus, monitoring is both informed by and guides control behaviours.

The process of solving complex problems

Fischer et al. (2012) have proposed an integrative theory on the process of complex problem solving. According to this theory, CPS involves the acquisition of implicit and/or explicit knowledge about relevant aspects of a problem, as well as the application of this knowledge in the search for a solution (compare with Funke, 2001; Fischer, 2015).

1. **Knowledge acquisition**: in order to sufficiently understand the problem, a problem solver has to systematically generate information (search for informative data), to sufficiently integrate this information into a viable mental model of the situation (search for adequate hypotheses), and to selectively focus on most relevant, central and urgent aspects of the problem (search for viable data-representations).
2. **Knowledge application**: based the explicit and implicit knowledge acquired, the problem solver has to make a set of interdependent decisions[3] (dynamic decision making) and to continuously monitor prerequisites and consequences of these decisions (monitoring) in order to systematically solve the problem at hand.

The interplay between knowledge acquisition and knowledge application is in a way similar to the idea of the generation and reduction of information, because an adequate subset has to be considered from all the knowledge available for effective knowledge application (for instance, structural knowledge has to be referred to only in cases when no sufficient instance knowledge is available). The advantage of the new formulation proposed by Fischer et al. (2012) is the broader perspective connected to the term "knowledge" with a lot of differentiation.

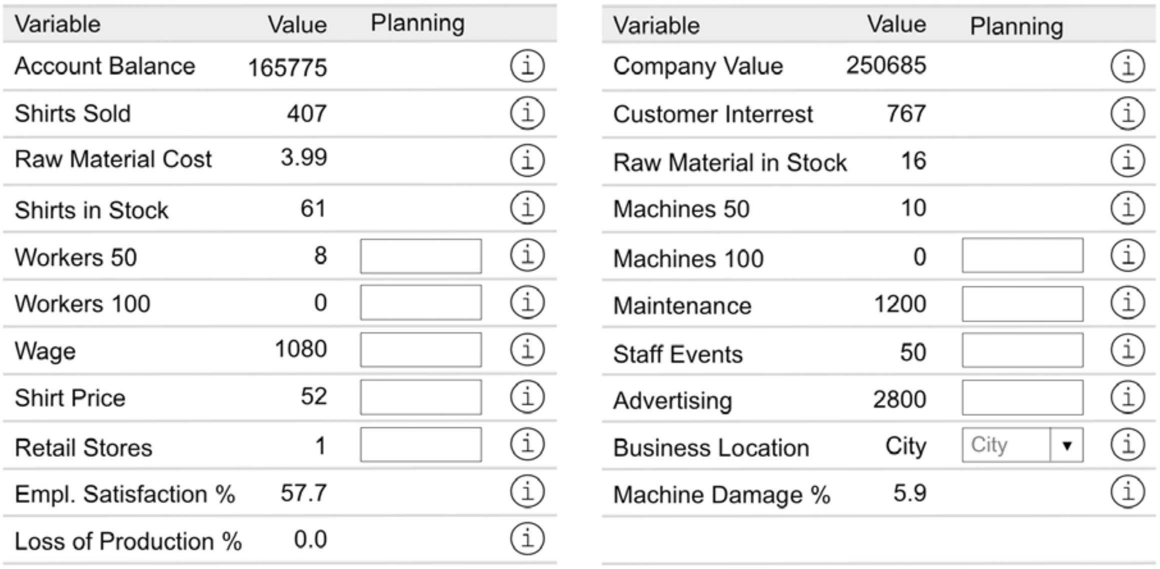
Assessment of complex problem solving

Dörner (1986) introduced the concept of operative intelligence as the crystalline ability to co-ordinate one's cognitive operations in a goal-oriented, wise and sustainable way: "Operative intelligence implies declarative and procedural knowledge about when and how to apply certain operation, as well as implicit knowledge about both the utility of applying and the contexts in which to remember operations" (compare with Fischer et al., 2012; Dörner and Schölkopf, 1991). As the various processes involved in CPS are intertwined and can hardly be assessed in isolation, Dörner (1986) proposed using simulations of complex problems to assess operative intelligence. Indicators for this kind of problem-solving competency in computer-based CPS simulations are typically based on states or processes of the system's variables (such as capital, population, or number of alarm messages, compare with Strohschneider and Güß, 1999), on the problem solver's inputs to the system (such as information requests or decisions made, compare with Dörner, 1986) or on separate knowledge tests that are administered afterwards or simultaneously (compare with Greiff, 2012).

During the 1990s the research community increasingly elaborated the prerequisites for a psychometric approach to CPS, that is, which kind of simulations to use and which scores to rely on. Some issues with complex problems had to be solved in order to use CPS simulations as reliable assessment instruments. First, they lacked comparability, as it proved hard to identify common aspects of different complex problems. Complex problems can be highly heterogenous in nature which hindered reliable scoring of specific CPS skills (Fischer, 2015). Funke (1992, 2001) proposed a solution to this problem, introducing formal frameworks to compare different complex problems based on their underlying causal structure. A series of experimental studies identified important structural determinants of a complex problem's difficulty: aspects such as the number of interrelations (Funke, 1985), the amount of dynamic changes (Funke, 1992) or the delay in feedback (Funke, 1985; Heineken, Arnold, Kopp and Soltysiak, 1992) were empirically proved to influence a complex problem's difficulty (for an overview, see Funke, 2003).

This introduction of formal frameworks for describing complex problems in a comparable manner was a first important prerequisite for the psychometric approach to CPS and a fruitful approach within CPS research (Funke, 2001). But even with the lack of comparability solved, in the 1990s many studies on CPS in assessment contexts were criticised for another methodological shortcoming that may have led to mixed results in early CPS research. When performance scores were built based on a scenario's states at multiple points in time (e.g., after every input), each state depended not only on the last input but on every prior input (i.e., measurement errors were not independent from each other).

Figure 7.2 **Screenshot of the Tailorshop problem**



Variable	Value	Planning	
Account Balance	165775		ⓘ
Shirts Sold	407		ⓘ
Raw Material Cost	3.99		ⓘ
Shirts in Stock	61		ⓘ
Workers 50	8		ⓘ
Workers 100	0		ⓘ
Wage	1080		ⓘ
Shirt Price	52		ⓘ
Retail Stores	1		ⓘ
Empl. Satisfaction %	57.7		ⓘ
Loss of Production %	0.0		ⓘ

Variable	Value	Planning	
Company Value	250685		ⓘ
Customer Interrest	767		ⓘ
Raw Material in Stock	16		ⓘ
Machines 50	10		ⓘ
Machines 100	0		ⓘ
Maintenance	1200		ⓘ
Staff Events	50		ⓘ
Advertising	2800		ⓘ
Business Location	City	City ▼	ⓘ
Machine Damage %	5.9		ⓘ

Source: Danner et al. (2011a), "Measuring performance in dynamic decision making: Reliability and validity of the Tailorshop simulation".

Take the Tailorshop problem (see Figure 7.2) for example: when a person is instructed to maximize company value, company value v after each month i could be measured as indicators (v_i) for the person's ability. But as company value changes by an additive amount c each month ($v_i = v_{i-1} + c_i$), the indicators are not experimentally independent and measurement errors would be correlated. In order to resolve the problem of correlated measurement errors, Danner and colleagues (2011a) proposed using the change in capital value after each month (c_i) – or the sign of this change (s_i) – as an adequate indicator of a person's competency. Results indicate that the sum of s_i over all but the first months was a highly reliable and valid estimate[4] of a person's problem-solving competency that was substantially correlated to performance on the Heidelberg finite state automaton (HEIFI; see Figure 7.3) and to Supervisor Ratings, even beyond reasoning (measured through Advanced Progressive Matrices). A tau-equivalent model (with factor loadings of all indicators on a single construct being constant and all measurement errors being constant as well as independent from each other) fitted the data sufficiently well (RMSEA = .07; CFI = .98). According to Danner et al. (2011b) a latent CPS factor (indicated by performance in the Tailorshop and the HEIFI simulation) was related to Supervisor Ratings even after controlling for IQ (measured by the Berlin Intelligence Structure Model test) whereas IQ did not have a unique contribution when compared to CPS.

A different approach to resolve the problem of correlated measurement errors is to apply a set of multiple complex systems (MCS approach; Greiff, Fischer et al., 2013). In contrast to single time-consuming simulations of highly complex problems, the idea is to use a larger number of less complex problems to allow for reliable scoring of specific problem solving skills (Funke, 2010). A typical example of this approach is the MicroDYN[5] test of Greiff, Wüstenberg et al. (2013) which consists of multiple independent simulations. Each simulation is based on a dynamic linear equation model with six variables (three independent variables that can be manipulated by the problem solver and three dependent variables that directly depend on the input variables). The problem solver has to find out about the system structure by interacting with the simulation (knowledge acquisition) and to reach certain goal-values for the dependent variables by a series of decisions concerning the independent variables (knowledge application). The time per simulation is limited to about 5 minutes and after each simulation a score for knowledge acquisition and knowledge application is derived. The simulations applied in the MicroDYN test are far less complex than the Tailorshop,[6] allowing reliable multi-item testing of specific problem-solving skills – the dual search for information and hypotheses, and the subsequent search for a multi-step solution (Greiff, Fischer et al., 2013) – even if it does not represent the full range of CPS requirements (Fischer, 2015). The skills reliably assessed by MicroDYN are important for many kinds of CPS beyond measures of reasoning (e.g. Greiff, Fischer et al., 2013; Wüstenberg, Greiff and Funke, 2012), but of course it takes more than these skills to solve complex problems of various kinds (Fischer, 2015). As Funke put it, “two measurement approaches, Tailorshop and MicroDYN, illustrate two different understandings of CPS. Whereas Tailorshop stands for a broad interpretation of CPS but has some weak points from a psychometric point of view, MicroDYN represents the psychometric sound realization of selected but important CPS aspects” (2010:138).

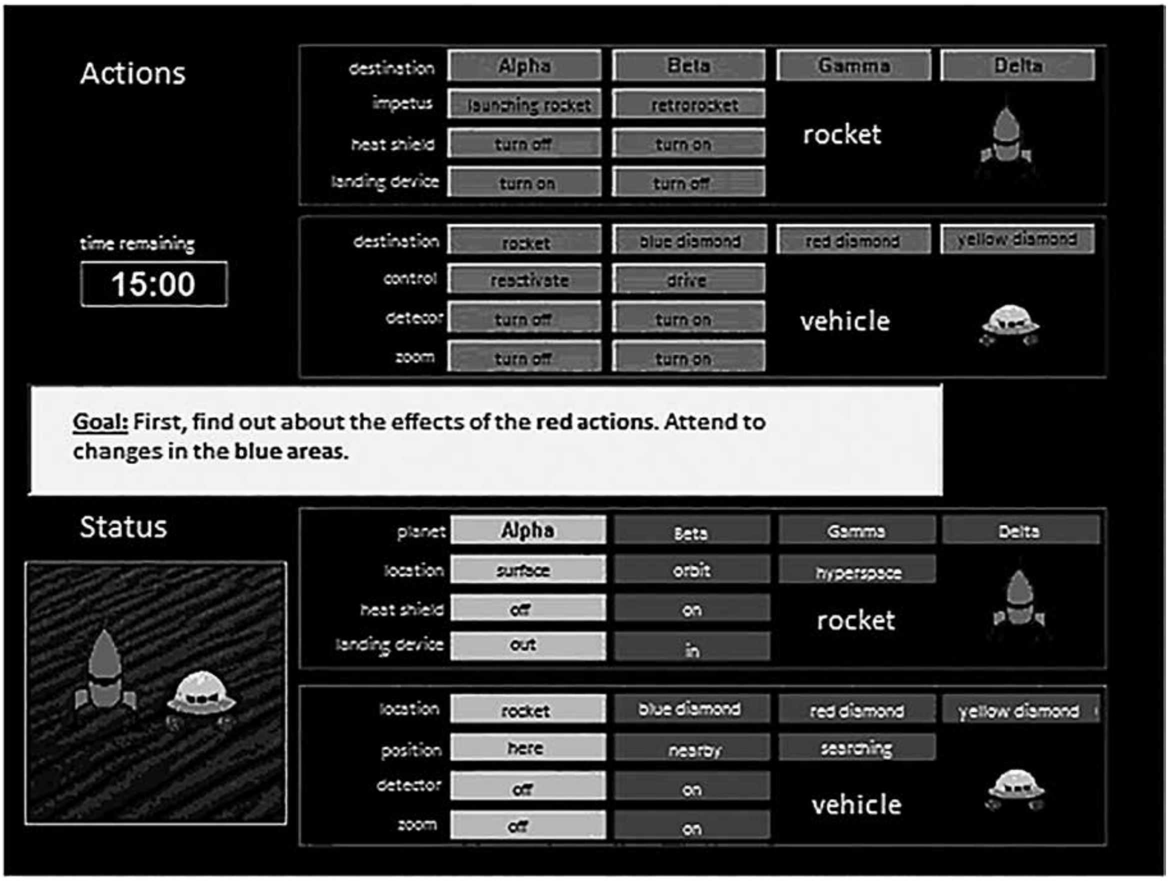
Discussion

Having considered the history of CPS we now review some of the more recent developments and outline possible trends for future research directions. These research directions are strongly aligned with PISA (OECD, 2010). The assessment of CPS in PISA is a showcase for an educationally motivated application of a construct originating from cognitive psychology and how this research can be meaningfully applied in the context of a large-scale assessment.

For the first time in a large-scale context, CPS skills were assessed in a national extension study to PISA 2000 in Germany. The finite state machine Space Shuttle was used to assess the skills related to interactive knowledge acquisition and knowledge application, as described above. Leutner and colleagues (2005) were able to identify a two-dimensional latent factor structure for Space Shuttle with one factor explaining the interactive knowledge acquisition in complex situations and another factor explaining knowledge application. Both aspects of CPS proved to be correlated to but separable from general intelligence[7], analytical problem solving (APS), and domain-specific competencies (Leutner et al., 2005).

These national efforts were followed by an assessment of analytical problem solving (APS) in the PISA 2003 international survey in all participating countries using a paper-and-pencil test. This kind of APS is closely related to the CPS skills measured by MicroDYN even after controlling for reasoning (Fischer et al., 2015).

Figure 7.3 **Screenshot of the finite state machine HEIFI**



Actions

destination	Alpha	Beta	Gamma	Delta
impetus	launching rocket	retrorocket	rocket	
heat shield	turn off	turn on		
landing device	turn on	turn off		

time remaining

15:00

destination	rocket	blue diamond	red diamond	yellow diamond
control	reactivate	drive	vehicle	
detecor	turn off	turn on		
zoom	turn off	turn on		

Goal: First, find out about the effects of the red actions. Attend to changes in the blue areas.

Status

planet	Alpha	Beta	Gamma	Delta
location	surface	orbit	hyperspace	
heat shield	off	on	rocket	
landing device	out	in		

location	rocket	blue diamond	red diamond	yellow diamond
position	here	nearby	searching	
detector	off	on	vehicle	
zoom	off	on		

Source: Wirth and Funke (2005), "Dynamisches Problemlösen: Entwicklung und Evaluation eines neuen Messverfahrens zum Steuern komplexer Systeme", p. 57.

The most recent PISA cycle in 2012 included MicroDYN and MicroFIN tasks to complement the static APS tasks as indicators of domain-general problem-solving competency. This was a major breakthrough for problem-solving research, because in PISA 2012, all participating countries could opt for computer-based assessment, and thus CPS could finally be included in the main survey. Besides the switch in mode of delivery, there were at least two more reasons why CPS was included in PISA 2012. First, PISA aims to capture skills relevant for real-life success and has made efforts to go beyond the mere assessment of skills in the classical literacy domains. To this end, Autor, Levy and Murnane (2003) remind us that over the last decades non-routine skills such as problem solving have gained in importance as routine skills decrease in our everyday lives. Thus, the OECD (2010) identifies problem solving as one of the key competencies relevant for acquiring domain-specific skills for success in life, rendering CPS a prime candidate for educational assessment and explaining its inclusion in PISA 2012. Second, psychometrically sound assessment is a prerequisite for any educational large-scale assessment and the unresolved assessment issues in CPS discussed above had delayed the inclusion of CPS in PISA at earlier stages. With the psychometric advances of the MicroDYN test, a valid and psychometrically acceptable assessment of important CPS skills became feasible and was, therefore, implemented in PISA 2012 as an innovative and first-time construct going beyond the classical literacy domains usually targeted in surveys such as PISA.

Trends for future research

Based on these considerations, two points can be concluded. First, the psychometric advances in the assessment of CPS as outlined above have equipped researchers with the means they need to better understand the cognitive structures underlying CPS and its relation to other constructs (e.g. Fischer, 2015). Second, assessment of CPS is not only theoretically important, but yields additional information shown by the empirically reported added value of CPS outlined above. To this end, with the results of PISA 2012 reporting how countries differ in their CPS level and how these differences may be related to other variables on different levels of analysis, CPS is highly relevant for educational research and extremely exciting for policy makers and educators around the world.

As we have seen, complex problems are manifold: usually there are many variables, multiple goals and dynamic dependencies. The interdependencies are often nonlinear, not transparent, and time-delayed, and sometimes there are random shocks, oscillations and interactions. It is difficult to combine all of these under the common roof of a standardised assessment instrument. However, each of the psychometric approaches described above manage to capture some of these aspects of complex problems and allows assessing non-routine problem-solving skills in the sense of Autor et al. (2003).

But are the psychometric advances reported so far already the end of the entire CPS story? Probably not. PISA 2015 merges social aspects with complex problem solving into a construct labelled “collaborative problem solving”. While in individual CPS, problem solvers have to interact with a dynamically changing system, collaborative problem solving extends this interaction towards exchange with other problem solvers. Obviously, working in teams and efficiently solving problems together constitutes a set of skills increasingly needed in the 21st century. However, an assessment of collaborative problem solving (see Chapter 14) faces many obstacles and has a long way to go. Considering success story told for CPS over recent years, this should encourage researchers to further pursue the path towards a valid and standardised assessment of collaborative problem solving.

Notes

[1] All these problems can be considered complex, as multiple highly interrelated aspects have to be considered in order to systematically transform a given situation into a desirable situation (Fischer, Greiff and Funke, 2012; Weaver, 1948). As Rey and Fischer pointed out,“Complexity refers to the number of interacting elements that have to be processed in working memory simultaneously in order to understand them” (2013:409).

[2] Please note, that this short review can by no means be complete: there are many other theories on cognitive processes that may also be considered relevant for CPS such as representational change (Ohlsson, 1992), and there are theories on the complex interactions of cognitive processes with motivational and emotional processes, such as Dörner et al. (2002). For a recent overview concerning theories on CPS please refer to Fischer et al. (2012).

[3] Either a solution is implemented instantly (triggered by the current goals, the situation perceived, and/or the content of working memory) or as a result of searching processes (such as planning, means-end analysis, hill-climbing, analogy, or domain-specific knowledge). If implementation fails or no solution can be derived at all, the problem solver is assumed to switch back to knowledge acquisition in order to change means or goals (Fischer et al., 2012).

[4] Please note, the sum of changes ci was less valid, a result that could partly be explained by outliers.

5 The MicroDYN test is based on multiple simulations based on dynamic linear equations. A similar test of the same skills, MicroFIN, is based on multiple simulations of finite state machines (Greiff, Fischer et al., 2013; Fischer, 2015).

6 In principle the MicroDYN approach would allow the implementation of problems with arbitrary complexity and difficulty. Up to now, research has focused on additive main effects of exogenous and endogenous variables, but the formal framework also allows for powers or interactions of variables to be included in models as additive terms (e.g. b*b or b*c).

7 In general, the relation between general intelligence and problem-solving performance seems to depend on the amount of prior knowledge in an inverted U-shaped way: correlations are low if there is either 1) not enough knowledge to solve the problem; or 2) sufficient knowledge to not have a problem at all (compare the Elshout-Raaheim-Hypothesis, see Leutner, 2002).

References

Autor, D.H., F. Levy and R.J. Murnane (2003), "The skill content of recent technological change: An empirical exploration", *Quarterly Journal of Economics*, Vol. 118/4, pp. 1279-1333.

Amelung, D. and J. Funke (2013), "Dealing with the uncertainties of climate engineering: Warnings from a psychological complex problem solving perspective", *Technology in Society*, Vol. 35/1, pp. 32-40, http://dx.doi.org/10.1016/j.techsoc.2013.03.001.

Berry, D.C.and D.E. Broadbent (1987), "The combination of explicit and implicit learning processes in task control", *Psychological Research*, Vol. 49/1,pp. 7-15.

Berry, D.C. and D.E. Broadbent (1984), "On the relationship between task performance and associated verbalizable knowledge", *Quarterly Journal of Experimental Psychology Section A*, Vol. 36/2, pp. 209-231.

Brehmer, B. (1992), "Dynamic decision making: Human control of complex systems", *Acta Psychologica*, Vol. 81/3, pp. 211-241.

Broadbent, D.E., P. Fitzgerald and M.H.P. Broadbent (1986), "Implicit and explicit knowledge in the control of complex systems", *British Journal of Psychology*, Vol. 77, pp. 33-50.

D'Zurilla T.J. and M.R. Goldfried (1971), "Problem solving and behavior modification", *Journal of Abnormal Psychology*, Vol.78/1, pp. 107-126.

Danner, D. et al. (2011a), "Measuring performance in dynamic decision making: Reliability and validity of the Tailorshop simulation", *Journal of Individual Differences*, Vol. 32, pp. 225-233, http://dx.doi.org/10.1027/1614-0001/a000055.

Danner, D. et al. (2011b), "Beyond IQ. A latent state-trait analysis of general intelligence, dynamic decision making, and implicit learning", *Intelligence*, Vol. 39/5, pp. 323-334, http://dx.doi.org/10.1016/j.intell.2011.06.004.

Dienes, Z. and R. Fahey (1995), "The role of specific instances in controlling a dynamic system", *Journal of Experimental Psychology: Learning, Memory and Cognition*, Vol. 21, pp. 848-862.

Dörner, D. (1997), *The Logic of Failure: Recognizing and Avoiding Error in Complex Situations*, Basic Books, New York.

Dörner, D. (1996), "Der Umgang mit Unbestimmtheit und Komplexität und der Gebrauch von Computersimulationen" [Coping with uncertainty and complexity and using computer simulations], in A. Diekmann and C.C. Jäger (eds.), *Umweltsoziologie: Sonderheft 36 der Kölner Zeitschrift für Soziologie und Sozialpsychologie*, pp. 489-515.

Dörner, D. (1989), *Die Logik des Misslingens: Strategisches Denken in komplexen Situationen* [The Logic of Failure: Strategic Thinking in Complex Situations], Reinbeck, Hamburg.

Dörner, D. (1986), "Diagnostik der operativen Intelligenz" [Assessment of operative intelligence], *Diagnostica*, Vol. 32/4, pp. 290-308.

Dörner, D. (1982), "Über die Schwierigkeiten menschlichen Umgangs mit Komplexität" [On human problems in coping with complexity], *Psychologische Rundschau*, Vol. 31, pp. 163-179.

Dörner, D. (1975). "Wie Menschen eine Welt verbessern wollten" (How people wanted to improve the world), *Bild der Wissenschaft*, Vol. 12/2, pp. 48-53.

Dörner, D., U. Drewes and F. Reither (1975), "Über das Problemlösen in sehr komplexen Realitätsbereichen" [On problem solving in very complex domains of reality], in W.H. Tack (ed.), *Bericht über den 29. Kongress der DGfPs in Salzburg 1974*, Vol. 1, Hogrefe, Göttingen: Hogrefe, pp. 339-340.

Dörner, D., H.W. Kreuzig, F. Reither and T. Stäudel (1983*), Lohhausen: Vom Umgang mit Unbestimmtheit un Komplexität* [Lohhausen: On Handling Complexity], Huber, Bern.

Dörner, D. and Schölkopf (1991), "Controlling complex systems; or, expertise as 'grandmother's know-how'", in K.A. Ericsson and J. Smith (eds.), *Towards a General Theory of Expertise: Prospects and Limits,*Cambridge University Press, New York, pp. 218-239.

Dörner, D. et al. (2002), *Die Mechanik des Seelenwagens* [Mechanics of the Soul Waggon], Huber, Bern.

Fischer, A. (2015), "Assessment of problem-solving skills by means of multiple complex systems – Validity of finite automata and linear dynamic systems", Dissertation theseis, Heidelberg University.

Fischer, A., S. Greiff and J. Funke (2012), "The process of solving complex problems", Journal of Problem Solving, Vol. 4/1, pp. 19-42, http://dx.doi.org/10.7771/1932-6246.1118.

Fischer, A., S. Greiff, S. Wüstenberg, J. Fleischer, F. Buchwald and J. Funke (2015), "Assessing analytic and interactive aspects of problem solving competency", *Learning and Individual Differences*, Vol. 39, pp. 172-179, http://dx.doi.org/10.1016/j.lindif.2015.02.008.

Frensch, P.A. and J. Funke (1995), "Definitions, traditions, and a general framework for understanding complex problem solving", in P.A. Frensch and J. Funke (eds.*), Complex Problem Solving: The European Perspective*, Erlbaum, Hillsdale, NJ, pp. 3-25.

Funke, J. (2010), "Complex problem solving: A case for complex cognition?", *Cognitive Processing*, Vol. 11/2, pp. 133-142, http://dx.doi.org/10.1007/s10339-009-0345-0.

Funke, J. (2003). *Problemlösendes Denken* [Problem Solving and Thinking], Kohlhammer, Stuttgart.

Funke, J. (2001), "Dynamic systems as tools for analysing human judgement", *Thinking and Reasoning*, Vol. 7/1, pp. 69-89, http://dx.doi.org/10.1080/13546780042000046.

Funke, J. (1992), *Wissen über dynamische Systeme: Erwerb, Repräsentation und Anwendung* [Knowledge about Dynamic Systems: Acquisition, Representation and Application], Springer, Heidelberg.

Funke, J. (1985), "Steuerung dynamischer Systeme durch Aufbau und Anwendung subjektiver Kausalmodelle" [Controlling dynamic systems by acquiring and applying subjective causal models], *Zeitschrift für Psychologie*, Vol. 193, pp. 435-457.

Gaschler, R. and P.A. Frensch (2007), "Is information reduction an item-specific or an item-general process?", *International Journal of Psychology*, Vol. 42/4, pp. 218-228.

Gonzalez, C., J.F. Lerch and C. Lebiere (2003), "Instance-based learning in dynamic decision making", *Cognitive Science*, Vol. 27/4, pp. 591-635, http://dx.doi.org/10.1016/S0364-0213(03)00031-4.

Gonzalez, C., P. Vanyukov and M.K. Martin (2005), "The use of microworlds to study dynamic decision making", *Computers in Human Behavior*, Vol. 21/2, pp. 273-286, http://dx.doi.org/10.1016/j.chb.2004.02.014.

Greiff, S. (2012), *Individualdiagnostik komplexer Problemlösefähigkeit* [Assessment of Complex Problem-Solving Ability], Waxmann, Munster.

Greiff, S., A. Fischer, S. Wüstenberg, P. Sonnleitner, M. Brunner and R. Martin (2013), "A multitrait–multimethod study of assessment instruments for complex problem solving", *Intelligence*, Vol. 41/5, pp. 579-596.

Greiff, S., S. Wüstenberg, G. Molnár, A. Fischer, J. Funke and B. Csapó (2013), "Complex problem solving in educational settings – Something beyond *g*: Concept, assessment, measurement invariance, and construct validity, *Journal of Educational Psychology*, Vol. 105/2, pp. 364-379, http://dx.doi.org/10.1037/a0031856.

Heineken, E., H.-J. Arnold, A. Kopp and R. Soltysiak (1992), "Strategien des Denkens bei der Regelung eines einfachen dynamischen Systems unter verschiedenen Totzeitbedingungen" [Strategies of thinking during regulation of simple dynamic systems with different conditions of reaction time], *Sprache & Kognition*, Vol. 11, pp. 136-148.

Hestenes, D. (1992), "Modeling games in the Newtonian world", *American Journal of Physics*, Vol. 60/8, pp. 732-748.

Jansson, A. (1994), "Pathologies in dynamic decision making: Consequences or precursors of failure?", *Sprache & Kognition*, Vol. 13/3, pp. 160-173.

Klahr, D. and K. Dunbar (1988), "Dual space search during scientific reasoning", *Cognitive Science*, Vol 12/1, pp. 1-48.

Klauer, K.C. (1993), *Belastung und Entlastung beim Problemlösen: Eine Theorie des deklarativen Vereinfachens* [Charge and discharge in problem solving: A theory of declarative simplification], Hogrefe, Göttingen.

Klein, G. (2008), "Naturalistic decision making", *Human Factors*, Vol. 50/3, pp. 456-460, http://dx.doi.org/10.1518/001872008X288385.

Klein, G., J. Orasanu, R. Calderwood and C.E. Zsambok (eds.) (1993), Decision Making in Action: Models and Methods", Ablex Publishing, Norwood, NJ.

Kluge, A. (2008), "What you train is what you get? Task requirements and training methods in complex problem-solving", *Computers in Human Behavior*, Vol. 24/2, pp. 248-308, http://dx.doi.org/10.1016/j.chb.2007.01.013.

Kretzschmar, A. and H.-M. Süß (2015), A study on the training of complex problem solving competence, *Journal of Dynamic Decision Making*, Vol. 1/4, http://dx.doi.org/10.11588/jddm.2015.1.15455.

Lantermann, E.-D., E. Döring-Seipel, B. Schmitz and P. Schima (2000), *SYRENE: Umwelt- und Systemlernen mit Multimedia*, Hogrefe, Göttingen.

Leutner, D. (2002), "The fuzzy relationship of intelligence and problem solving in computer simulations", *Computers in Human Behavior*, Vol. 18/6, pp. 685-697, http://dx.doi.org/10.1016/S0747-5632(02)00024-9.

Leutner, D., J. Wirth, E. Klieme and J. Funke (2005), "Ansätze zur Operationalisierung und deren Erprobung im Feldtest zu PISA 2000" [Approaches to operationalizations and their testing in the PISA 2000 field trial], in E. Klieme, D. Leutner and J. Wirth (eds.), *Problemlösekompetenz von Schülerinnen und Schülern*, VS Verlag für Sozialwissenschaften, Wiesbaden, pp. 21-36.

Liebeck, H. (2008), "Problemlösetraining" [Training of problem solving], in M. Linden and M. Hautzinger (eds.), *Verhaltenstherapiemanual,*Springer, Heidelberg, pp. 203-207.

Meadows, D., D. Meadows, E. Zahn and P. Milling (1972), *Die Grenzen des Wachstums: Bericht des Club of Rome zur Lage der Menschheit* [The Limits of Growth: Report of the Club of Rome Concerning the Human Situation], Deutsche Verlags-Anstalt, Stuttgart.

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

OECD (2010), *PISA 2012 Field Trial Problem Solving Framework*, OECD, Paris.

Ohlsson, S. (1992), "Information processing explanations of insight and related phenomena", In M.T. Keane and K.J. Gilhooly (eds.), *Advances in the Psychology of Thinking*, Vol. 1, Harvester Wheatsheaf, London, pp. 1-44.

Osman, M. (2010), "Controlling uncertainty: A review of human behavior in complex dynamic environments", *Psychological Bulletin*, Vol. 136/1, pp. 65-86, http://dx.doi.org/10.1037/a0017815.

Putz-Osterloh, W. (1981), "Über die Beziehung zwischen Testintelligenz und Problemlöseerfolg" [On the relation between test intelligence and problem solving performance], *Zeitschrift für Psychologie*, Vol. 189, pp. 79-100.

Rey, G.D., and A. Fischer (2013), "The expertise reversal effect concerning instructional explanations", *Instructional Science*, Vol 41/2, pp. 407-429, http://dx.doi.org/10.1007/s11251-012-9237-2.

Schaub, H. (2006), "Störungen und Fehler beim Denken und Problemlösen", in J. Funke (ed.), *Denken und Problemlösen,* Enzyklopädie der Psychologie, Themenbereich C: Theorie und Forschung, Serie II: Kognition, Band 8, Hogrefe, Göttingen, pp. 447-482.

Schoppek, W. (2002), "Examples, rules, and strategies in the control of dynamic systems", *Cognitive Science Quarterly*, Vol. 2, pp. 63-92.

Schunn, C.D. and D. Klahr (2002), "Multiple-space search in a more complex discovery microworld", in D. Klahr (ed.), *Exploring Science: The Cognition and Development of Discovery Processes*, Cambridge, MA, pp. 161-199.

Simon, H.A. and G. Lea (1974), "Problem solving and rule induction: A unified view", in L.W. Gregg (ed.), *Knowledge and Cognition*, Erlbaum, Hillsdale, NJ, pp. 105-127.

Strohschneider, S. and D. Güß (1999), "The fate of the Moros: A cross-cultural exploration of strategies in complex and dynamic decision making", *International Journal of Psychology*, Vol. 35/4, pp. 235-252.

Vester, F. (2007), *The Art of Interconnected Thinking: Tools and Concepts for a New Approach to Tackling Complexity*, MCB, Munich.

Vollmeyer, R., B.D. Burns and K.J. Holyoak (1996), "The impact of goal specifity and systematicity of strategies on the acquisition of problem structure", *Cognitive Science*, Vol. 20/1, pp. 75-100.

Weaver, W. (1948), "Science and complexity", *American Scientist*, Vol. 36/4, pp. 536-544.

Wirth, J. and J. Funke (2005), "Dynamisches Problemlösen: Entwicklung und Evaluation eines neuen Messverfahrens zum Steuern komplexer Systeme" [Dynamic problem solving: Development and evaluation of a new device for the control of complex systems], in E. Klieme, D. Leutner and J. Wirth (eds.), *Problemlösekompetenz von Schülerinnen und Schülern*, VS Verlag für Sozialwissenschaften, Wiesbaden, pp. 55-72.

Wittmann, W.W. and H.-M. Süß (1999), "Investigating the paths between working memory, intelligence, knowledge, and complex problem solving performances via Brunswik symmetry", in P.L. Ackerman, P.C. Kyllonen and R.D. Roberts (eds.), *Learning and Individual Differences: Process, Trait, and Content Determinants*, American Psychological Association, Washington, DC, pp. 77-108.

Wüstenberg, S., S. Greiff and J. Funke (2012), "Complex problem solving: More than reasoning?", *Intelligence*, Vol. 40/1, pp. 1-14, http://dx.doi.org/10.1016/j.intell.2011.11.003.

PART III

Empirical results

Chapter 8

Empirical study of computer-based assessment of domain-general complex problem-solving skills

By

Gyöngyvér Molnár

MTA-SZTE Research Group on the Development of Competencies, University of Szeged, Hungary.

Samuel Greiff

Maison des Sciences Humaines, Université du Luxembourg, Luxembourg.

Sascha Wüstenberg

TWT GmbH Science & Innovation, Germany.

and Andreas Fischer

Forschungsinstitut Betriebliche Bildung (f-bb), Nuremberg, Germany.

This study reviews the results of a recent project on problem solving. Taking a developmental and structural perspective, it contrasts static, paper-and-pencil tests with interactive, technology-based tests of thinking skills, with a special reference to reasoning skills including knowledge acquisition, knowledge application, and transfer of knowledge. Hungarian students aged 11 to 17 completed problem-solving tests in static scenarios (assessing domain-specific problem-solving skills from maths and science) and in interactive scenarios (assessing domain-general complex problem-solving skills). The students were also assessed for inductive reasoning and fluid intelligence, both by first generation tests. This chapter uses the results to elicit evidence for the development of dynamic problem solving and for the relationship between those 21st-century skills. Finally, it discusses the possibility of using traditional static tests to predict performance in third generation tests measuring dynamic problem solving.

This research was funded by the European Union (Project ID number: TÁMOP–3.1.9–11/1–2012–0001), the SZTE-MTA Research Group on the Development of Competencies, the K115497 OTKA research program, and by grants from the Fonds National de la Recherche Luxembourg (ATTRACT "ASKI21") and the German Federal Ministry of Education and Research (BMBF FKZ 01JG1062). – Some of the analyses and figures as well as minor text parts presented in this article are reiterations of previous work and are taken directly from there (for instance, from Molnár, Greiff, & Csapó, 2013). At the same time, much of the theoretical thoughts and the analyses are extensions and are unique contributions of this chapter and have not been published elsewhere. Concerning terminology, please note that there exist different labels for the current subject (e.g., complex problem solving, dynamic problem solving, creative problem solving). In this chapter we use the prefix "dynamic".

Introduction

How can we assess knowledge and skills essential in the 21st century? Can we predict students' performance in so-called "third generation" tests that measure, for instance, dynamic problem solving skills, from their performance in traditional, static and more academic testing situations (so-called "first generation" tests)? How do problem-solving skills develop over time, especially the skills involved in coping with interactive and dynamically changing problems? Are students ready for third generation testing or do they prefer first and second generation tests?

The assessment of 21st-century skills has to provide students with an opportunity to demonstrate their skills related to the acquisition and application of knowledge in new and unknown problem situations. These are the skills needed in today's society, characterised as it is by rapid change, where the nature of applicable knowledge changes frequently and specific content becomes quickly outdated (de Koning, 2000). There are three ways to assess 21st-century skills. First, they can be measured through traditional approaches using first generation tests – with designs based closely on existing static paper-and-pencil tests, such as the typical tests of domain-specific problem-solving-skills. Second, they can be measured with second generation tests using new formats, including multimedia, constructed response, automatic item generation and automatic scoring tests (Pachler et al., 2010). Finally, they can be assessed through third generation tests which allow students to interact with complex simulations and dynamically changing items, dramatically increasing the number of ways they can demonstrate skills such as the domain-general dynamic problem solving-skills in the study presented below. All three approaches are relevant to avoid methodological artefacts, since measurement of a thinking skill always involves reference to general mental processes independent of the context used (Ericson and Hastie, 1994).

In this chapter we take a developmental and a structural perspective to elaborate on technology-based assessment, contrasting static, paper-and-pencil tests with interactive, technology-based tests of thinking skills, with special reference to reasoning skills including knowledge acquisition, knowledge application and transfer of knowledge. We synthesise research related to identifying 21st-century skills, including these skills, both in static tests assessing domain-specific problem solving (DSPS) and in interactive scenarios assessing domain-general dynamic problem solving (DPS). We elicit evidence for the development of these domain-general 21st-century skills measured by a third generation test, and for the relationship between those skills and more domain-specific static problem-solving skills measured by first generation tests, inductive reasoning (IR) and fluid intelligence, which is considered a "hallmark" indicator of the general *g* factor (intelligence), both of which are also measured by first generation tests. Finally, we discuss the possibility of predicting performance in third generation tests of DPS from performance in traditional, static testing situations.

Technology-based assessment and new areas of educational assessment

Information and communication technologies have fundamentally changed the possibilities and the process of educational assessment. Research and development in technology-based assessment (TBA) go back three decades. In the 1990s the focus was on exploring the applicability of a broad range of technologies from the most common, widely available computers to the most expensive, cutting-edge technologies for assessment purposes (Baker and Mayer, 1999). A decade later, large-scale international assessments were conducted to explore the potential and the implementation of TBA, such as the National Assessment of Educational Progress (NAEP), the Programme for International Student Assessment (PISA), and the Progress in International Reading Literacy Study (Csapó et al., 2012; Fraillon et al., 2013; OECD, 2010, 2014a) with the aim of replacing traditional paper-and-pencil test delivery with assessments exploiting the manifold advantages of computer-based delivery.

The predominant subject of these studies was to compare the results of paper-and-pencil and computer-based assessments of the same construct (Kingston, 2009; Wang et al., 2008). The instruments studied were mostly first and third generation tests (Pachler et al., 2010). As technology has become more ubiquitous over the past decades, familiarity with computers (Mayrath, Clarke-Midura and Robinson, 2012) and test mode effects should no longer be much of an issue (Way, Davis and Fitzpatrick, 2006). In the last five years computer-based assessment has opened doors to exploring features not yet studied, such as domain-general dynamic problem-solving skills. This required the development of third generation tests that did not rely on standard, multiple-choice item formats (Ripley et al., 2009; Greiff, Wüstenberg and Funke, 2012; Greiff et al., 2014; OECD, 2014b). Finally, a relatively recent development is the online assessment of group processes in place of individual assessments, for example projects by the Assessment and Teaching of 21st Century Skills (ATC21S); Griffin, McGaw and Care (2012); PISA; National Assessment and Testing; Hesse et al. (2015); and the OECD (2013).

From static to dynamic problem solving with reference to reasoning skills

Reasoning is one of the most general thinking skills (Pellegrino and Glaser, 1982; Ropo, 1987) and often understood as a generalised capability to acquire, apply and transfer knowledge. It is related to, and the strongest component of, almost all higher-order cognitive skills and processes (Csapó, 1997), such as general intelligence (Klauer and Phye, 2008) and problem solving (Gentner, 1989; Klauer, 1996). In connection with reasoning, problem-solving skills have been extensively studied from different perspectives over the past decade, as they involve the ability to acquire and use new knowledge, or to use pre-existing knowledge in order to solve novel (i.e. non-routine) problems (Sternberg, 1994). Studies into the different approaches have showed the need to distinguish between domain-specific and domain-general problem-solving skills (Sternberg, 1995).

One of the arguably most comprehensive international large-scale assessments, the PISA survey, places special emphasis on DSPS and DPS processes. It measured DSPS in 2003 and DPS in 2012 as an additional domain beyond the usual fields of reading, science and mathematics (OECD, 2005, 2010, 2013; Greiff, Holt and Funke, 2013; Fischer et al., 2015). Here, problem solving was seen as a "cross-curricular" domain.

In DSPS situations, problem solvers need to combine knowledge acquired in and out of the classroom to reach the desired solution by retrieving and applying previously acquired knowledge in a specific domain. In this study, we treat DSPS as a process of applying domain-specific – mathematical and scientific – knowledge in three different types of new situations: 1) complete problems, where all necessary information to solve the problem is given at the outset; 2) incomplete problems relying on missing information that students are expected to have learnt at school; and 3) incomplete problems relying on missing information that was not learnt at school (see Molnár, Greiff and Csapó, 2013).

In contrast, DPS tasks require an additional series of complex cognitive operations (Funke, 2010; Raven, 2000) beyond those involved in DSPS tasks (e.g., Klieme, 2004; Wüstenberg, Greiff and Funke, 2012). In DPS tasks, students have to directly interact with a problem situation which is dynamically changing over time and use the feedback provided by the computer to acquire and apply new knowledge (Fischer, Greiff and Funke, 2012; Funke, 2014). This measures the competencies of knowledge acquisition and knowledge application.

Recent analyses provide evidence for the relation between domain-specific static and domain general-dynamic problem solving, especially between the acquisition and application of academic and non-academic knowledge; and their relation to general mental abilities.

Aims

The objective of the study described here is to review the results of a recent project on DPS (see Greiff et al., 2013; Molnár, Greiff and Csapó, 2013a, 2013b; Wüstenberg et al., 2012). We thus intend to answer five research questions:

1) Can DPS be better modelled as a two-dimensional construct with knowledge acquisition and knowledge application as separate factors than as a one-dimensional construct with both dimensions subsumed under one factor?

2) Can DPS be shown to be invariant across different grades implying that differences in DPS performance can be validly interpreted?

3) How does DPS develop over time during public education in different school types?

4) What is the relationship between DPS, DSPS and IR and do these relationships change over time?

5) Can we predict achievement in DPS (measured by a third generation test) by performance in more traditional testing situations (DSPS, IR and intelligence) assessing thinking skills including problem solving?

Methods

Sample

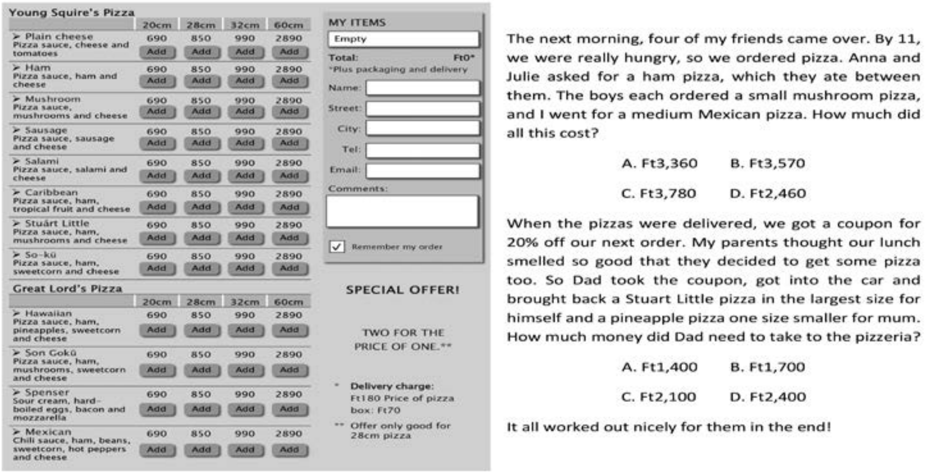
The samples of the study were drawn from 5th to 11th grade students (aged 11 to 17) in Hungarian primary and secondary schools. There were 300 to 400 students in each cohort. One-third of the secondary school students (9th to 11th grade) were from grammar schools, with the rest from vocational schools. Some technical problems occurred during online testing resulting in completely random data loss. Participants who had more than 50% of data missing (316 students) were excluded from the analyses. The final sample for the analyses contained data from 788 students.[1]

Instruments

The instruments of the study include tests of DPS, DSPS, IR, fluid reasoning (an important indicator of intelligence) and a background questionnaire. One comprehensive version of the DPS test was used regardless of grade. The test consisted of seven tasks created following the MicroDYN approach (see Greiff, et al., 2013; Chapter 6). In the first stage, participants were provided with instructions including a trial task. Subsequently, students had to explore several unfamiliar systems, in which they had to find out how variables were interconnected and draw their conclusions in a situational model (knowledge acquisition; Funke, 2001). In the final stage, they had to control the system by reaching given target values (knowledge application; see Greiff et al., 2013).

Three versions of the DSPS test were used with different levels of item difficulty, which varied by school grade. The test versions contained anchor items allowing performance scores to be represented on a single scale in each case. Approximately 80% of the 54 DSPS items were in multiple-choice format, the remaining items were constructed-response questions and all of the problems presented information in realistic formats. The DSPS test comprised three types of problems: 1) problems where all the information needed to solve the problem was given at the outset (knowledge application); 2) incomplete problems requiring the use of additional information previously learnt at school as part of the National Core Curriculum (knowledge application and knowledge transfer); and finally 3) incomplete problems requiring the use of additional information that had not been learnt at school and needed to be retrieved from real-life knowledge (knowledge application, knowledge transfer and knowledge creation; see Molnár et al., 2013a). The presentation of the DSPS problems looked similar in all three versions of the test. The left-hand column presented information in realistic formats (such as a map, picture or drawing) and on the right was a story of a family trip or a class excursion and a prompt students to solve problems (e.g., using the information provided and supplementing it with school knowledge) as they would arise during the trip" (Figure 8.1). All of the problems needed domain-specific knowledge from the field of mathematics or science to solve.

Figure 8.1 **Example of tasks in the domain-specific problem-solving test**



Young Squire's Pizza

	20cm	28cm	32cm	60cm
➢ Plain cheese Pizza sauce, cheese and tomatoes	690 Add	850 Add	990 Add	2890 Add
➢ Ham Pizza sauce, ham and cheese	690 Add	850 Add	990 Add	2890 Add
➢ Mushroom Pizza sauce, mushrooms and cheese	690 Add	850 Add	990 Add	2890 Add
➢ Sausage Pizza sauce, sausage and cheese	690 Add	850 Add	990 Add	2890 Add
➢ Salami Pizza sauce, salami and cheese	690 Add	850 Add	990 Add	2890 Add
➢ Caribbean Pizza sauce, ham, tropical fruit and cheese	690 Add	850 Add	990 Add	2890 Add
➢ Stuárt Little Pizza sauce, ham, mushrooms and cheese	690 Add	850 Add	990 Add	2890 Add
➢ So-kü Pizza sauce, ham, sweetcorn and cheese	690 Add	850 Add	990 Add	2890 Add

Great Lord's Pizza

	20cm	28cm	32cm	60cm
➢ Hawaiian Pizza sauce, ham, pineapples, sweetcorn and cheese	690 Add	850 Add	990 Add	2890 Add
➢ Son Gokü Pizza sauce, ham, mushrooms, sweetcorn and cheese	690 Add	850 Add	990 Add	2890 Add
➢ Spenser Sour cream, hard-boiled eggs, bacon and mozzarella	690 Add	850 Add	990 Add	2890 Add
➢ Mexican Chili sauce, ham, beans, sweetcorn, hot peppers and cheese	690 Add	850 Add	990 Add	2890 Add

MY ITEMS

Empty

Total: Ft0*

*Plus packaging and delivery

Name:

Street:

City:

Tel:

Email:

Comments:

☑ Remember my order

SPECIAL OFFER!

TWO FOR THE PRICE OF ONE.**

* Delivery charge: Ft180 Price of pizza box: Ft70

** Offer only good for 28cm pizza

The next morning, four of my friends came over. By 11, we were really hungry, so we ordered pizza. Anna and Julie asked for a ham pizza, which they ate between them. The boys each ordered a small mushroom pizza, and I went for a medium Mexican pizza. How much did all this cost?

A. Ft3,360 B. Ft3,570

C. Ft3,780 D. Ft2,460

When the pizzas were delivered, we got a coupon for 20% off our next order. My parents thought our lunch smelled so good that they decided to get some pizza too. So Dad took the coupon, got into the car and brought back a Stuart Little pizza in the largest size for himself and a pineapple pizza one size smaller for mum. How much money did Dad need to take to the pizzeria?

A. Ft1,400 B. Ft1,700

C. Ft2,100 D. Ft2,400

It all worked out nicely for them in the end!

Source: Molnár et al. (2013a), "Inductive reasoning, domain specific and complex problem solving: Relations and development".

The IR test comprised both open-ended and multiple-choice items. It was divided into three subtests: number analogies (14 items) and number series (16 items) embedded in mathematical contexts, and verbal analogies (28 items; see Csapó, 1997; Figure 8.2).

The Culture Fair Test 20-R (CFT) test was used for measuring students' fluid intelligence, which is according to state-of-the-art intelligence theories such as the Cattell–Horn–Carrolltheory one of the most important indicators and markers of g (Weiß, 2006), the core of intelligence (Carroll, 2003). It consisted of 4 subscales with 56 figural items.

The background questionnaire contained questions regarding students' socio-economic background, academic achievement, school subject attitudes and parental education.

Figure 8.2 **Example of tasks in the inductive reasoning test**

Verbal analogies	CHAIR : FURNITURE = DOG : ? a CAT b ANIMAL C SPANIEL d TABLE e DOGHOUSE
Number analogies	20→32 :: 8→20 :: 11→____

Note: The original items were in Hungarian.

Source: Molnár et al. (2013a), "Inductive reasoning, domain specific and complex problem solving: Relations and development".

Procedure

The tests were completed in four sessions, each lasting approximately 45 minutes. In Session 1, students worked on the DPS test. In Session 2 students had to complete the DSPS test, in Session 3 they completed the intelligence test (CFT) and in Session 4 an IR test and the background questionnaire.

The online data collection was carried out by means of the TAO platform over the Internet. The testing took place in the computer labs of the participating schools, using existing computers and preinstalled browsers.

Confirmatory factor analyses (CFA) were applied to test the underlying measurement model of DPS (research question 1). A weighted least squares mean and variance adjusted (WLSMV) estimator and theta parameterisation were used to estimate model parameters, since all items were scored dichotomously (Muthén and Muthén, 2010). Tucker-Lewis Index (TLI), comparative fit index (CFI), and root mean square error of approximation (RMSEA) were proposed to assist in determining model fit (see Vandenberg and Lance, 2000). Nested model comparisons were conducted using the DIFFTEST procedure (Muthén and Muthén, 2010). All measurement models were computed with Mplus (Muthén and Muthén, 2010).

Measurement invariance was tested by multigroup analyses (MACS) within the structural equation modelling (SEM) approach (research question 2). The testing procedure for categorical data involves a fixed sequence of model comparisons (Vandenberg and Lance, 2000), testing three different levels of invariance (configural invariance, strong factorial invariance and strict factorial invariance) by comparing measurement models from the least to the most restrictive model (for more detail see Greiff et al., 2013).

Rasch's model was used for scaling the data, and then linear transformation of the logit metric was chosen. The means of 8th graders were set to 500 with a standard deviation of 100 (research question 3). A four-parameter logistic equation was used for the curve fitting procedures to estimate development. A coefficient of determination (R^2) was computed to express how well the model described the data (see Molnár et al., 2013a). SEM was also used to examine the direct and indirect effects between DPS, DSPS, IR and intelligence (research questions 4 and 5).

Results

Descriptive statistics

Internal consistencies across the DPS, DSPS, IR and CFT tests were generally high (DPSα = .92; DSPS$_{levels1to3}$ α = .73, .82, .65; IRα = .95; CFTα = .88, respectively; see Molnár et. al., 2013a and Greiff et al., 2013). However, there was a noticeable drop in reliability in the DSPS test used in grades 9 to 11 (Level 3) down to .65. Grade-level analyses reveal increased probability of measurement error among 9th and 10th grade students, whereas the reliability of the same test among 11th grade students (DSPS$_{grade11}$$\alpha$ = .72) proved to be higher (see Molnár et al., 2013a). For this reason, we included only the data for 5th to 8th grade and 11th grade students in all further analyses of DSPS.

Dimensionality of dynamic problem solving

The two-dimensional model of DPS measured by MicroDYN includes knowledge acquisition and knowledge application as separate factors. This was compared to a one-dimensional model that combined both dimensions under one general factor. In accord with our theoretical hypothesis, a two-dimensional model of DPS showed a better fit than a one-dimensional one (χ^2-difference test: χ^2 = 86.121; df = 1; p < .001). Fit indices were significantly better in the two-dimensional model (Table 8.1; see Greiff et al., 2013). In summary, this provided empirical support for the theoretically derived two-dimensional model of knowledge acquisition and knowledge application in DPS.

Table 8.1 **Goodness of fit indices for testing dimensionality of the dynamic problem solving model**

Model	χ^2	df	p	CFI	TLI	RMSEA
2 dimensional model	164.06	53	.001	.967	.978	.050
1 dimensional model	329.35	52	.001	.912	.944	.079

Note: df = degrees of freedom; CFI = comparative fit index; RMSEA = root mean square error of approximation; TLI = Tucker-Lewis Index.

Measurement invariance of the domain-general dynamic problem-solving instrument

In order to test measurement invariance of DPS, we tested configural invariance, strong factorial invariance and strict factorial invariance and compared the measurement models. All measurement models yielded a good fit as indicated by CFI, TLI and RMSEA (Table 8.2; CFI, TLI > .95; RMSEA < .06).

Neither the strong factorial invariance nor the strict factorial invariance models differed significantly from the configural invariance model, implying that measurement invariance was maintained (ΔCFI < .01 and non-significant χ^2-difference test, Table 2).

Table 8.2 **Goodness of fit indices for measurement invariance of DPS in the MicroDYN approach**

Model	χ^2	df	$\Delta\chi^2$	Δdf	p	CFI	TLI	RMSEA
Configural inv.	161.04	104	–	–	–	.975	.975	.051
Strong fact.inv.	170.10	115	22.29	23	>.10	.976	.982	.047
Strict fact.inv.	165.82	116	53.15	43	>.10	.978	.983	.045

Note: fact.inv. = factorial invariance; df = degrees of freedom; CFI = comparative fit index; RMSEA = root mean square error of approximation; TLI = Tucker-Lewis Index. Nested model comparisons (computing $\Delta\chi^2$) were conducted using the DIFFTEST procedure, a special $\Delta\chi^2$ difference test (see Muthén and Muthén, 2010).

As the DPS was measured as invariant, mean differences across groups attending different grades could be interpreted as true differences in the underlying DPS skills and were not due to psychometrical issues (Byrne and Stewart, 2006). This allowed us to make valid direct comparisons of skill levels between students and between classes and investigate the development of DPS in research question 3.

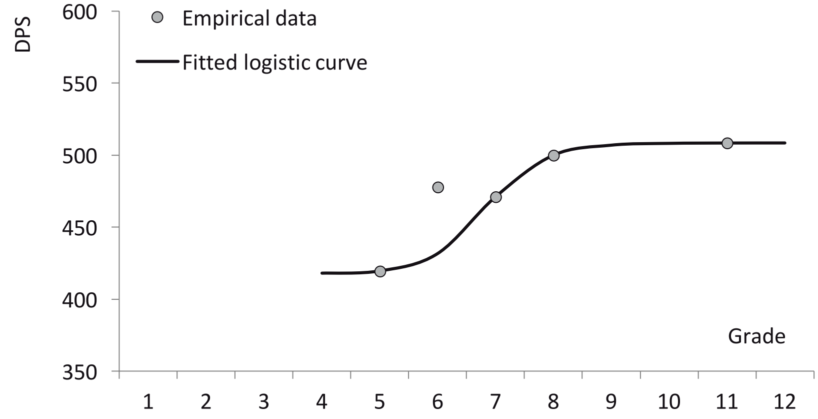
Development of domain-general dynamic problem solving

The features of developmental tendencies of DPS were in line with the findings of previous studies regarding thinking skills, (e.g., Adey et al., 2007) such as problem solving (Molnár et al., 2013a), and follow a regular developmental trend. The development spans several years and can be described with a logistic curve that fitted the empirical data adequately ($R^2 = .91$; see Figure 8.3). The fit was perfect ($R^2 = 1.00$) if we excluded 6th grade students from the analyses. The behaviour of 6th grade students calls for further study because the currently available empirical data do not account for this phenomenon.

The fastest growth, the point of inflexion, was observed in 7th grade (at the age of 12.8), offering opportunities for training in DPS. This is the sensitive period for stimulation, since the enhancement of thinking skills is most effective when “students’ development is still in progress, especially when they are in a fast-growing phase” (Molnár et al., 2013a). All in all, elementary school students from 5th to 8th grade showed noticeable development in DPS; fostering DPS could be very efficient during this period. This trend seemed to change after 8th grade, when development slowed down. However, extrapolation of the fitted logistic curves indicated that substantial development took place before 3rd grade and some improvement can also be expected after 11th grade (Figure 8.3).

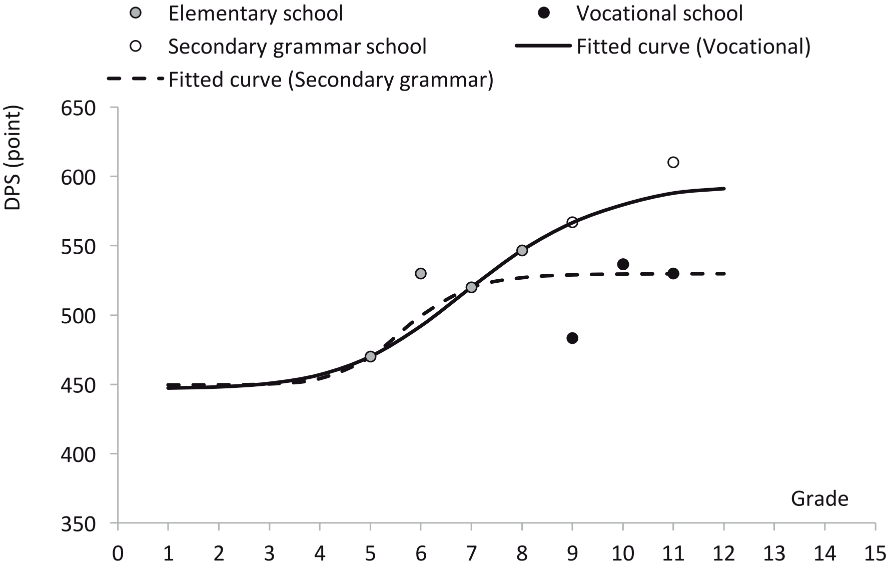
Looking at the analyses of the different school types, the coefficient of determination decreased in the case of vocational school (R^2 = .72) and increased (R^2 = .96) for grammar school data. The slope and maximum of the developmental curve changed notably after primary school. Even the 9th graders in grammar schools performed better than 11th graders in vocational schools, and 11th graders in vocational schools performed worse than 8th graders in elementary schools. The performance differences (t = -8.59, p < .01) between vocational and grammar school students proved to be stable over time (Figure 8.4; Molnár et al., 2013b).

Figure 8.3 **Developmental curve of dynamic problem solving**



Source: Molnár, G., S. Greiff and B. Csapó (2013b), "Relations between problem solving, intelligence and socio-economic background", paper presented at the EARLI, Munich, 27-31 August 2013.

Figure 8.4 **Developmental curve of dynamic problem solving by school type**



Source: Molnár, G., S. Greiff and B. Csapó (2013b), "Relations between problem solving, intelligence and socio-economic background", paper presented at the EARLI, Munich, 27-31 August 2013.

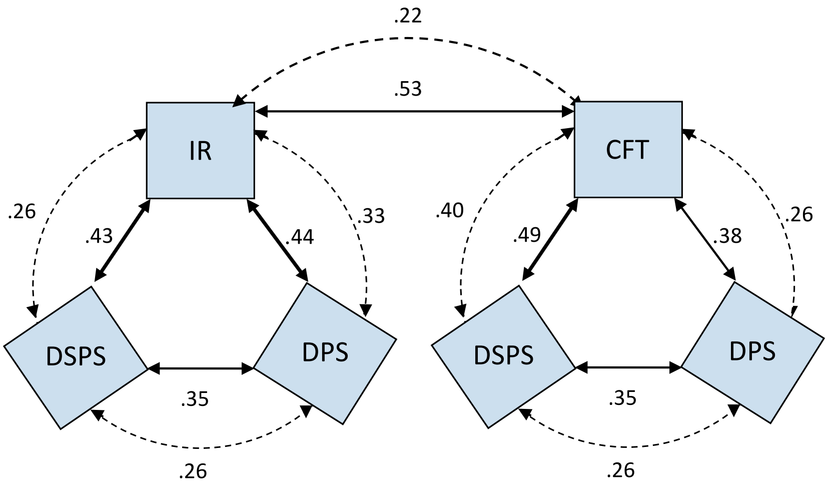
The relationship between inductive reasoning, intelligence, domain-specific and domain-general dynamic problem solving

The bivariate correlations between inductive reasoning, DSPS and DPS were similar to those between intelligence, DSPS and DPS. All of them were moderate ranging from .35 to .49 (Figure 8.5). The relationships proved to be similar between IR and either DSPS or DPS (r = .43 and .44, p < .01, respectively) and between intelligence and DSPS (r = .49, p < .01). They were significantly stronger (z = 1.80, p < .05) than the correlation between DSPS and DPS (r = .35, p < .01) or between intelligence and DPS (r = .38, p < .01).The relationship between IR and intelligence, measuring inductive reasoning skills (e.g. generalisation, discrimination) in academic and in figural content, respectively, proved to be the strongest (r = .53, p < .01).

Partial correlations were significantly lower as all bivariate relationships were influenced by the third construct, that is, either DPS; DSPS, or reasoning/intelligence (r_{ir_DSPS} = .26; r_{ir_DSPS} = .33; r_{DSPS_DPS} =.26, r_{CFT_DSPS} = .40; r_{CFT_DPS} = .26; p < .01 respectively).

To analyse the stability of these processes, three different cohorts were selected and analysed separately: 5th graders, whose skills showed some development, but who had not yet entered the fast-growing phase, 7th graders, who were in the fast-growing developmental phase regarding DPS and 11th graders, who had left behind the fast-growing phase and were approaching the end of compulsory education. The correlation patterns showed instability across the three cohorts as they proved to be more homogeneous within grades than across grades. The strengths of the correlations of any two constructs were not generally influenced by the third construct (see Molnár et al., 2013a).

Figure 8.5 **Correlations between inductive reasoning, intelligence, domain-specific and domain-general problem solving**



Note: All coefficients are significant at p < .01; partial correlations are depicted as dotted lines. IR: inductive reasoning; CFT: culture-fair test of intelligence; DSPS: domain-specific problem solving; DPS: domain-general problem solving.

Source: Molnár et al. (2013a).

On the whole, the correlation between DPS and IR was the highest and the most stable over time. It was not due to students' level of DSPS skills. A possible explanation for this phenomenon could be that the basic mechanisms of IR, such as finding similarities and dissimilarities, or generating rules based on observation, are also the basic cognitive processes involved in DPS.

The relationship between DPS and DSPS became stronger (from no significant correlation to moderate but significant correlation) over time. This can be explained by the fact that the strategies used in DSPS and DPS situations become more similar, the mechanisms involved getting closer, over time. Older students have more opportunities to acquire and apply knowledge in a self-regulated way so DPS may be a prerequisite to DSPS. DSPS – if it is embedded in a domain-specific context as in our study – is based on knowledge application, whereas DPS – as captured in the MicroDYN approach – is a prerequisite to gaining and applying new knowledge.

The predictive power of first generation tests measuring thinking skills on DPS

The analysis regarding DSPS and DPS indicated a moderate correlational relationship between performance in static domain-specific and performance in domain-general dynamic problem solving environments (see research question 4). Thus, we could conclude that more traditional first generation tests of problem solving (DSPS) could predict performance in third generation tests of problem solving (DPS) and the other way around in a synchronised manner. To analyse the causal and one-way predictive force of first generation tests on third generation tests, SEM analyses are needed.

In the first, simplest SEM model, DPS measured by a third generation test was regressed on DSPS assessed by a first generation static test. The measurement model showed a good fit for the overall sample (CFI = 1.00, TLI = 1.00, RMSEA = .00). The standardised path coefficient was β = .33. A significant amount of variance remained unexplained.

In the second model, DSPS and CFT were used as predictors for DPS. The measurement model still showed a good fit for the overall sample (CFI = 1.00, TLI = 1.00, RMSEA = .00). The standardised path coefficients were .19 for DSPS and .29 for CFT/intelligence. That is, DSPS predicted performance in DPS beyond CFT. However, a significant amount of variance still remained unexplained and path coefficients were only moderately high.

In the third model, DPS was regressed on DSPS, CFT and IR. The standardised path coefficients dropped (to .13, .18 and .26 respectively), indicating the role of IR in the predictive power of DPS. The fit of the measurement model was good. A significant amount of variance remained unexplained in this model as well. Accordingly – knowing the limitations of the analyses, namely that although the tests measure problem solving, they do not measure the same construct in a strict sense (see above) – it is possible to predict performance in dynamic problem solving, from performance in traditional, static tests of DSPS, IR or intelligence, but the power of prediction is limited.

Thus, third generation tests are designed to require more cognitive skills and therefore additional aspects of problem solving that are relevant in today's life but are not captured by classical first generation tests of domain-specific skills..

Discussion

Our aim was to review the results of a recently conducted project that investigated different thinking skills and their relevance in educational settings (see Greiff et al., 2013; Molnár et al., 2013a, 2013b). We concentrated on dynamic problem solving, which we approached from an assessment perspective. We analysed its dimensionality and measurement invariance, and tested how it is influenced by other cognitive constructs such as domain-specific problem solving, fluid intelligence and reasoning assessed by domain-specific problem solving, the Culture Fair Test and inductive reasoning tests in an educational context.

Generally, the results of the study provided support for a view of DPS as a set of indispensable skills for the 21st century having relevance to educational settings from a developmental perspective. Our analysis shows that DPS can be better understood as involving two factors, knowledge acquisition and knowledge application, rather than subsuming both dimensions under one factor. DPS proved to be measurement invariant across different school grades, implying that differences in DPS performance can be validly interpreted. DPS developed following a regular trend, spanning several years of compulsory schooling, and the development curve varied across different school types. The correlations between DPS, DSPS, and IR were moderate and changed over time, as did those between DPS, DSPS and intelligence. With some limitations, it was possible to predict performance in dynamic problem solving measured by a third generation test by performance in traditional, static testing situations assessing thinking skills including problem solving (see also Fischer et al., 2015).

Dimensionality

More specifically, the MicroDYN approach gave students the opportunity to demonstrate their level of skills regarding both acquiring and applying knowledge. This confirmed previous research results reporting substantial but not perfect correlations between knowledge acquisition and knowledge application (e.g. Bühner et al., 2008; Kröner et al., 2005; Wüstenberg et al., 2012) regarding problem solving, which was identified by Mayer and Wittrock (2006) and confirmed to be significant by PISA (OECD, 2010, 2014b). Our result corroborated the hypothesis that knowledge acquisition is a necessary but not a sufficient condition of knowledge application (Greiff et al., 2013). These are two general and overarching processes of problem representation and solution at a broad level. However, one could object that knowledge acquisition and its application are themselves composed of several secondary processes or skills, each of which may be relevant in educational settings, and that those were neglected in the current study (see Greiff et al., 2013). Future research should address this question and define the components of lower-level processes.

Invariance

The measurement of these two dimensions was invariant across the students tested, from 5th to 11th grade. That means individual differences in DPS scores can be interpreted as true differences in the construct allowing direct comparisons of students in different grades and building a DPS scale from 5th grade to 11th grade. Such measurement invariance is a prerequisite for analyses on developmental patterns. However, this does not mean that the two factors of DPS, knowledge acquisition and application, relate to each other in the same way across all grades. Further research is needed to describe the changes in the relationship between knowledge acquisition and knowledge application with respect to DPS across different school grades and into adulthood.

Development

The development of DPS took place mostly during the compulsory schooling years, offering an opportunity to explicitly foster this higher-order thinking skill. As the greatest development was observed between 6th and 8th grade, that is the most sensitive and effective period in which to enhance students' DPS skills (Molnár et al., 2013a). This supports previous research results reporting relatively slow development of thinking skills (see, e.g., Csapó, 1997), and a lack of explicit training (de Koning, 2000; Molnár, 2011). Development occurs spontaneously as a "by-product" of schooling rather than as a result of explicit training (de Koning et al., 2002; Molnár et al., 2013a). Our results also confirm the importance of the elementary school years in the fostering of thinking skills as a possible way of making education more effective (Adey et al., 2007). To achieve this aim in practice, explicit training (Molnár, 2011) or different teaching methods (e.g. Shayer and Adey, 2002) have been suggested and proved to be effective tools.

The developmental curves differed notably after 8th grade in the different school types in the Hungarian school system. Mean performance differences between classes of vocational and grammar school students could be expressed in several developmental years and proved to remain stable over time. This confirmed previous research results showing high performance differences even in non-curriculum-related situations such as DPS, and reporting a drop in performance and motivation in 9th grade test scores, creating a gap between Hungarian vocational and grammar school students as a result of the strong selection process after primary school (OECD, 2005; Csapó, Molnár and Kinyó, 2008). Our findings support the claim that the effect of the extreme selectiveness exercised in curriculum-related situations, i.e. when students need to use knowledge acquired at school in order to find the desired solutions, can be detected in more general problem-solving situations as well, when domain-general cognitive processes are needed instead of content knowledge and rote learning (Buchner, 1995; Molnár et al., 2013a). Our results indicated a hypothetical significant role of IR and intelligence in the development of DSPS and DPS.

Relationships between thinking skills

The role of IR and intelligence in the development of DSPS and DPS proved to be significant, indicating that IR and intelligence are correlated with, but yet distinct from, the knowledge acquisition and application phase of DPS. As a result of the analyses, it can be assumed that interventions affecting one skill may affect the other as well. For instance, if DPS is fostered, DSPS and IR will be developed indirectly as well. This result draws attention to the importance of developing knowledge acquisition and knowledge application skills that also go along with improved reasoning skills (Hamers, De Koning and Sijtsma, 2000; Klauer, 1996), suggesting that together with IR (de Koning, 2000; Resnick, 1987), DPS should become an integral part of school agendas and should be incorporated into a broad range of school-related learning activities. This highlights the importance of developing of special methods to enhance DPS, such as guided discovery, or using DPS tasks as assessment and eventual enhancement tools to foster students' IR and domain-specific knowledge acquisition and application skills (see Molnár et al., 2013a).

The fact that strength of the relationships between the four thinking skills (IR, intelligence, DSPS and DPS) changed over time signals that good problem solvers somehow learn to know how to solve both DPS and DSPS tasks, or that problem solving strategies used in DSPS situations become more and more similar to the strategies used in domain general DPS problems (see also Fischer et al., 2015). This further emphasises the importance of explicit training of DPS. After a while the same mechanisms, the same knowledge acquisition and application skills are made use of when solving different kinds of problems (Molnár et al., 2013a), while the role of experience and knowledge in specific academic areas for solving problems decreases over time.

Predictive power of first generation tests

DPS, as measured with a state-of-the-art third generation test, was predicted by DSPS and intelligence measured by more traditional first generation tests. This confirms the results in research question 4 in general, while some variance remained unexplained. This may point towards the explanation for research question 4 and the results in this study and in the literature, specifically that DPS measures skills not measured by DSPS or intelligence, and predicts performance beyond reasoning in school grades (Wüstenberg et al., 2012). The results highlight once again the importance of DPS and, as a by-product, the importance of third generation testing, since DPS cannot be captured by first generation tests.

The present study contributes to the issue of assessing 21st-century skills by using third generation tests of cross-curricular skills indispensable for successful participation in modern Western society, thus overcoming some of the limitations associated with understanding only specific academic abilities such as reading, writing, and maths. We conclude by emphasising 1) the possibility of using and measuring new constructs by third generation tests; and 2) the potential of thinking skills such as dynamic problem solving as educationally relevant constructs that can and should be fostered over several years during public education. Encouraging the development of knowledge acquisition and knowledge application, that is, enhancing thinking skills relevant for both academic and non-academic contexts, is one of the most important and most challenging educational goals lying ahead.

Notes

[1] The same data were used as in Molnár et al. (2013a) and Greiff et al. (2013) but unique analyses were conducted here.

References

Adey, P., B. Csapó, A. Demetriou, J. Hautamaki, and M. Shayer (2007), "Can we be intelligent about intelligence? Why education needs the concept of plastic general ability", *Educational Research Review*, Vol. 2/2, pp. 75-97, http://dx.doi.org/10.1016/j.edurev.2007.05.001.

Baker, E.L. and R.E. Mayer (1999), "Computer-based assessment of problem solving", *Computers in Human Behavior*, Vol. 15/3-4, pp. 269-282.

Buchner, A. (1995), "Basic topics and approaches to the study of complex problem solving", in P.A. Frensch and J. Funke (eds.), *Complex Problem Solving: The European Perspective*, Erlbaum, Hillsdale, NJ, pp. 27 63.

Bühner, M., S. Kröner and M. Ziegler (2008), "Working memory, visual-spatial intelligence and their relationship to problem-solving" *Intelligence*, Vol. 364, pp. 672-680.

Byrne, B.M. and S.M. Stewart (2006), "The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process", *Structural Equation Modeling*, Vol. 13/2, pp. 287-321.

Carroll, J.B. (2003), "The higher-stratum structure of cognitive abilities: Current evidence supports *g* and about ten broad factors" in H. Nyborg (ed.), *The Scientific Study of General Intelligence: Tribute to Arthur R. Jensen*, Pergamon, Amsterdam, pp. 5-21.

Csapó, B. (1997), "The development of inductive reasoning: Cross-sectional assessments in an educational context", *International Journal of Behavioral Development*, Vol. 20/4, pp. 609-626.

Csapó, B., G. Molnár and L. Kinyó (2008), "Analysis of the selectiveness of the Hungarian educational system in international context", paper presented at the 3rd IEA International Research Conference in Taipai, Taiwan, 16-20 September 2008.

Csapó, B., J. Ainley, R. Bennett, T. Latour and N. Law (2012), "Technological issues of computer-based assessment of 21st century skills" in P. Griffin, B. McGaw and E. Care (eds.), *Assessment and Teaching of 21st Century Skills*, Springer, New York, pp. 143-230.

De Koning, E. (2000), *Inductive Reasoning in Primary Education: Measurement, Teaching, Transfer*, Kerckebosch, Zeist.

De Koning, E., J.H.M. Hamers, K. Sijtsma and A. Vermeer (2002), "Teaching inductive reasoning in primary education", *Developmental Review,* Vol. 22, pp. 211–241.

Ericsson, K.A. and R. Hastie (1994), "Contemporary approaches to the study of thinking and problem solving", in R.J. Sternberg (ed.), *Thinking and Problem Solving*, 2nd edition, Academic Press, New York, pp. 37-79.

Fischer, A., S. Greiff and J. Funke (2012), "The process of solving complex problems", *Journal of Problem Solving*, Vol. 4/1, pp. 19-42.

Fischer, A., S. Greiff, S. Wüstenberg, J. Fleischer, F. Buchwald, and J. Funke (2015), "Assessing analytic and interactive aspects of problem solving competency", *Learning and Individual Differences*, Vol. 39, pp. 172-179, http://dx.doi.org/10.1016/j.lindif.2015.02.008.

Fraillon, J., et al. (2013), "Measuring computer and information literacy across countries", paper prepared for the 5th IEA International Research Conference, Singapore, 26-28 June 2013, www.iea.nl/fileadmin/user_upload/IRC/IRC_2013/Papers/IRC-2013_Fraillon_etal.pdf.

Funke, J. (2014), "Analysis of minimal complex systems and complex problem solving require different forms of causal cognition", Frontiers in Psychology, Vol. 5, http://dx.doi.org/10.3389/fpsyg.2014.00739.

Funke, J. (2010), "Complex problem solving: A case for complex cognition?", *Cognitive Processing*, Vol. 11/2, pp. 133-142, http://dx.doi.org/10.1007/s10339-009-0345-0.

Funke, J. (2001), "Dynamic systems as tools for analysing human judgement", *Thinking and Reasoning*, Vol. 7/1, pp. 69-89, http://dx.doi.org/10.1080/13546780042000046.

Gentner, D. (1989), "The mechanisms of analogical learning", in S. Vosniadou and A. Ortony (eds.), *Similarity and Analogical Reasoning*, Cambridge University Press, London. pp. 199-241.

Greiff, S. (2012), "Assessment and theory in complex problem solving – A continuing contradiction?", *Journal of Educational and Developmental Psychology*, Vol. 2/1, pp. 49-56.

Greiff, S. and A. Fischer (2013), "Der Nutzen einer komplexen Problemlösekompetenz: Theoretische Überlegungen und Empirische Befunde" (On the usefulness of complex problem solving competency: Theoretical considerations and empirical results), *Zeitschrift für Pädagogische Psychologie*, Vol. 27/1, pp. 27-39.

Greiff, S., D.V. Holt and J. Funke (2013), "Perspectives on problem solving in cognitive research and educational assessment: Analytical, interactive, and collaborative problem solving", *Journal of Problem Solving*, Vol. 5, pp. 71-91, http://dx.doi.org/10.7771/1932-6246.1153.

Greiff, S., et al. (2014), "Domain-general problem solving skills and education in the 21st century", *Educational Research Review*, Vol. 13, pp. 74-83.

Greiff, S., S. Wüstenberg, and J. Funke (2012), "Dynamic problem solving: A new measurement perspective", *Applied Psychological Measurement*, Vol. 36/3, pp. 189-213.

Greiff, S., et al. (2013), "Complex problem solving in educational settings – Something beyond *g*: Concept, assessment, measurement invariance, and construct validity, *Journal of Educational Psychology*, Vol. 105/2, pp. 364-379, http://dx.doi.org/10.1037/a0031856.

Griffin, P., B. McGaw and E. Care (eds.) (2012), *Assessment and Teaching of 21st Century Skills*, Springer, Dordrecht.

Hamers, J.H.M., E. De Koning and K. Sijtsma (2000), "Inductive reasoning in third grade: Intervention promises and constraints", *Contemporary Educational Psychology*, Vol. 23/2, pp. 132-148.

Hesse, et al. (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* (pp. 37-56). Springer, Dordrecht.

Kingston N. M. (2009), "Comparability of computer- and paper-administered multiple-choice tests for K–12 populations: A synthesis", *Applied Measurement in Education*, Vol. 22/1, pp. 22-37.

Klauer, K.J. (1996), "Begünstigt induktives Denken das Lösen komplexer Probleme? Experimentellen Studien zu Leutners Sahel-Problem" (Has inductive thinking a positive effect on complex problem solving?), *Zeitschrift für Experimentelle Psychologie*, Vol. 43/1, pp. 361-366.

Klauer, K.J. and G.D. Phye (2008), "Inductive reasoning: A training approach", *Review of Educational Research*, Vol. 78/1, pp. 85-123.

Klieme, E. (2004), “Assessment of cross-curricular problem-solving competencies” in J.H. Moskowitz and M. Stephens (eds.), *Comparing Learning Outcomes: International Assessments and Education Policy*, Routledge Falmer, London, pp. 81–107.

Kröner, S., J.L. Plass and D. Leutner (2005), “Intelligence assessment with computer simulations” *Intelligence*, Vol. 33/4, pp. 347-368, http://dx.doi.org/10.1016/j.intell.2005.03.002.

Mayer, R.E. and M.C. Wittrock (2006), “Problem solving”, in P. A. Alexander and P. H. Winne (eds.), *Handbook of Educational Psychology*, Erlbaum, Mahwah NJ, pp. 287-303.

Mayrath, M., J. Clarke-Midura and D. Robinson (eds.) (2012), *Technology-Based Assessment for 21st Century Skills: Theoretical and Practical Implications from Modern Research*, Springer-Verlag, New York.

Molnár, G. (2011), “Playful fostering of 6- to 8-year-old students’ inductive reasoning”, *Thinking Skills and Creativity*, Vol. 6/2, pp. 91-99.

Molnár, G., S. Greiff and B. Csapó (2013a), “Inductive reasoning, domain specific and complex problem solving: Relations and development”, *Thinking Skills and Creativity*, Vol. 9/8, pp. 35-45, http://dx.doi.org/10.1016/j.tsc.2013.03.002.

Molnár, G., S. Greiff and B. Csapó (2013b), “Relations between problem solving, intelligence and socio-economic background”, paper presented at the EARLI, Munich, 27-31 August 2013.

Muthén, L.K. and B.O. Muthén (2010), *Mplus User’s Guide*, Muthén & Muthén, Los Angeles.

OECD (2014a), *PISA 2012 Results: What Students Known and Can Do (Volume 1, Revised Edition, February 2014): Student Performance in Mathematics, Reading and Science,* OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264208780-en.

OECD (2014b), *PISA 2012 Results: Creative Problem Solving (Volume V): Students’ Skills in Tackling Real-Life Problems*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264208070-en.

OECD (2013), *PISA 2015: Draft Collaborative Problem Solving Framework*, OECD, Paris, www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf.

OECD (2010), *PISA 2012 Field Trial Problem Solving Framework*, OECD, Paris.

OECD (2005), *Problem Solving for Tomorrow’s World: First Measures of Cross-Curricular Competencies from PISA 2003*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264006430-en.

Pachler, N., et al. (2010), “Formative e-assessment: Practitioner cases” *Computers & Education*, Vol. 54/3, pp. 715-721.

Pellegrino, J.W. and R. Glaser (1982), “Analyzing aptitudes for learning: Inductive reasoning” in R. Glaser (ed.), *Advances in Instructional Psychology*, Vol. 2. Erlbaum, Hillsdale, NJ, pp. 269-345.

Raven, J. (2000), “Psychometrics, cognitive ability, and occupational performance”, *Review of Psychology*, Vol. 7, pp. 51-74.

Resnick, L. (1987), *Education and Learning to Think*, National Academy Press, Washington, DC.

Ripley, M., et al. (2009), *Review of Advanced e-Assessment Techniques (RAeAT) Final Report*, Joint Information Systems Committee.

Ropo, E. (1987), “Skills for learning: A review of studies on inductive reasoning”, *Scandinavian Journal of Educational Research*, Vol. 31/1, pp. 31-39.

Shayer, M. and P. Adey (Eds.) (2002), *Learning Intelligence: Cognitive Acceleration Across the Curriculum from 5 to 15 Years*, Open University Press, Milton Keynes.

Sternberg, R.J. (1995), "Expertise in complex problem solving: A comparison of alternative concepts" in P.A. Frensch and J. Funke (eds.) (1995), *Complex Problem Solving: The European Perspective*, Erlbaum Associates, Hillsdale, NJ, pp. 295-321.

Sternberg, R. (ed.) (1994), *Thinking and Problem Solving*, Academic Press, San Diego.

Vandenberg, R.J and C.E. Lance (2000), "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research", *Organizational Research Methods*, Vol 3/1, pp. 4-70

Wang, S., et al. (2008), "Comparability of computer-based and paper-and-pencil testing in K–12 reading assessment: A meta-analysis of testing mode effects", *Educational and Psychological Measurement*, Vol. 68/1, pp. 5-24.

Way, W.D., L.L. Davis and S. Fitzpatrick (2006), "Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills", paper presented at the Annual Meeting of the National Council on Measurement in Education. San Francisco, April 2006.

Weiß, R. H. (2006), *Grundintelligenztest Skala 2 – Revision – (CFT 20-R)*, Hogrefe, Göttingen.

Wüstenberg, S., S. Greiff and J. Funke (2012), "Complex problem solving. More than reasoning?", *Intelligence*, Vol. 40/, pp. 1-14, http://dx.doi.org/10.1016/j.intell.2011.11.003.

Chapter 9

Factors that influence the difficulty of problem-solving items

By
Ray Philpot, Dara Ramalingam
Australian Council for Educational Research, Melbourne, Australia.
John A. Dossey
Distinguished Professor of Mathematics Emeritus, Illinois State University, United States.
and Barry McCrae
University of Melbourne and Australian Council for Educational Research, Australia.

This chapter presents a study undertaken by the authors using data from the PISA 2012 computer-based assessment of problem solving. It considered ten characteristics understood to influence the difficulty of items used in problem-solving assessement. Each item was rated to reflect the amount of each characteristic it possessed. The item responses from about 85 000 participants from 44 countries were analysed to obtain item response theory (IRT) estimates of item difficulties. The predictor characteristics were analysed in a number of ways, including a hierarchical cluster analysis, regression analysis with item difficulty as outcome variable and a principal component factor analysis. The main characteristics predicting difficulty seem to be: the complexity and type of reasoning skills involved in solving the problem; the amount of opportunity the solver is given to experiment or uncover hidden facets in a problem scenario (more opportunity to explore and experiment will make a problem easier); and the number and nature of constraints that a solution must satisfy (complex or conflicting constraints will make a problem more difficult).

Introduction

Individuals have different levels of ability in problem solving and so will find a given problem more or less difficult. Using Rasch measurement (Rasch, 1960), however, response data can be used to determine the difficulty of each of a set of assessment items (tasks) and place them on a scale. This difficulty measure of an item will be independent of the capability of anyone attempting the problem.

If the task characteristics of an item that contribute to its difficulty can be identified and quantified, one can then explore empirically which factors have the most influence on task difficulty. This in turn will enable the construction of items to be better targeted when assessing the problem-solving abilities of people in educational testing, personnel selection or psychological research.

This chapter outlines a study the authors conducted to determine the relationship between the characteristics of an item and the corresponding Rasch item difficulties in the PISA 2012 problem-solving assessment. The study informed the development of the final PISA 2012 problem-solving cognitive framework.

In this study, the term "item" refers to the material presented in the assessment that participants respond to – perhaps correctly, perhaps not – and the particular psychometric parameters (including "difficulty") that can be attributed to the item based on an analysis of the responses. Each item has a problem-solving "task" that the participants must engage with; the "task characteristics" consist of the cognitive demands and structural factors believed to affect the difficulty of the item.

Similar work examining factors that influence item difficulty has been carried out for reading and mathematics items in PISA. The interested reader is referred to Lumley et al. (2012) and Turner et al. (2012), respectively, for details.

Characteristics that might influence task difficulty

What characteristics or attributes of a problem-solving task might influence its difficulty? An example of a task characteristic that could plausibly influence the task's difficulty is the amount of information in a problem scenario that an individual has to engage with to successfully solve a problem. The more information that has to be taken into account in order to solve a problem, the more difficult the problem is likely to be. Indeed, large amounts of information may "overcharge" human processing capabilities (Fischer et al., 2012).

Apart from the sheer amount of information, the form in which it is presented might have an effect on difficulty: if the solver has to integrate two or more representations or understand an unfamiliar representation this may increase the cognitive load. Examples of different representations include text, tables, pictures, networks, lists and graphs.

The degree of familiarity a solver has with the problem situation itself (or the representations used) might affect the difficulty of an item, as the activation of prior knowledge and feelings of competency are well known to contribute to problem-solving proficiency (Mayer, 1992, 1998).

Regardless of the degree of familiarity a solver has with a particular scenario, if he or she tackles several tasks based on a single scenario, then later tasks are likely to be easier since the solver will have learned something about the system in attempting earlier tasks.

In many real-world problems, some information that is essential for solving a task might not at first be evident. A person confronted with such a problem may need to actively explore the environment or problem scenario to uncover the information needed to solve a task. For instance, the task might be to operate an unfamiliar piece of equipment without any instructions for how to do so. How easy it is to discover relevant information and the form of such information will affect the difficulty of the item.

A problem situation may have a more or less complex system or structure in terms of its underlying states and variables and in the relationships between its elements. For example, two MP3 players might look very similar (e.g. have the same number and positioning of buttons), but one might have far greater functionality than the other, or the sequence of button presses required to perform a particular action might be longer on one of the MP3 players than the other. In general, tasks based on problem situations with a high level of system complexity are likely to be harder than those based on situations with lower levels of complexity, all other things being equal.

Problem solving can be thought of as moving from an initial (given) state to a goal state (solution) using a series of operations in a problem space (e.g. Newell and Simon, 1972). The number of steps or operations that are needed to reach the solution may affect the difficulty of the task.

In most cases, the decision about which step to take at a given stage in a problem task will require the exercise of reasoning. The likely reasoning skills required to solve a problem can affect its difficulty, since some skills are more sophisticated than others. If sustained reasoning, or more than one type of reasoning is needed to reach a solution, the item will likely be more difficult.

A solution will usually be constrained in some way: either the task will contain explicit conditions to be satisfied or there will be implicit constraints present in the problem space. Constraints may even be conflicting. The number and type of constraints will affect the difficulty of the item.

In the computer-based problems in PISA 2012, test takers had scope for some degree of experimentation or could try out possibilities on screen, since the tasks were designed to be interactive. They could click buttons, navigate through pathways, change the values of the variables controlling the features of a device and so on, to see the effect these actions had. Exploring a (simulated) environment may assist the solver in developing a solution to a problem. Scenarios in which there is less opportunity for experimentation may therefore make the task more difficult, all other things being equal.

Table 9.1 summarises the above discussion, listing all these task characteristics together with a provisional qualitative assessment of each one's likely effect on a problem-solving item's difficulty. This list was first developed for the *PISA 2012 Problem Solving Framework* (OECD, 2013) and then subsequently refined by the authors.

Interactive versus static problem situations

The problems in the PISA 2012 problem-solving assessment are divided by the nature of the problem situation into two classes, interactive and static, as defined below. Each of the problems consists of an opening explanation of a context, or situation, followed by one or more tasks concerning the problem situation requiring a response from participating students. In all there were 16 problem situations divided into 42 tasks. Six of the problems consisted of two tasks each and ten of the problems consisted of three tasks each.

More than half of the problems in the PISA 2012 problem-solving assessment are interactive problems (see Chapter 5). These require solvers to explore the problem situation to acquire additional knowledge needed to solve the problem. Examples of interactive problems encountered in everyday life include discovering how to use an unfamiliar mobile telephone or automatic vending machine. These are examples of what in the research literature are often called complex problems (Frensch and Funke, 1995). Frensch and Funke define complex problems as follows:

The given state, goal state, and barriers between given state and goal state are complex, change dynamically during problem solving, and are intransparent. The exact properties of the given state, goal state, and barriers are unknown to the solver at the outset (French and Funke, 1995: 18).

The rest of the problems are known as static problems. These are problems in which all the information necessary to solve the problem is disclosed to the problem solver at the outset: they are completely transparent by definition. The Tower of Hanoi problem (Ewert and Lambert, 1932) is a classic example of a static problem. Note that all the logical consequences of the information might not be apparent to the solver, but all the data and the rules for manipulating it *are* available to the solver.

Using this vocabulary, interactive problems are intransparent (i.e. there is undisclosed information), but not necessarily dynamic or very complicated. For a discussion of the rationale behind using the term "interactive" rather than "complex", see Chapter 5.

Table 9.1. **Proposed task characteristics affecting item difficulty**

Characteristic	**Hypothesised effect on an item's difficulty**
Amount of information	The more information that is presented in a problem and the more complex it is, the more difficult the task is likely to be. (The information may be in the stimulus and/or the task statement.)
Representation of information	Unfamiliar representations, and multiple representations (especially when information presented in different representations has to be integrated), tend to increase item difficulty. (Includes text, diagrams, images, graphs, tables etc.)
Non-disclosure of information	The more that relevant information has to be discovered (e.g. effect of operations, automatic changes occurring in the system, unanticipated obstacles), the more difficult the item is likely to be.
System complexity	The difficulty of an item is likely to increase as the number of components or elements in the system increases and as they become more interrelated.
Constraints to be satisfied	Items where the task has a few simple constraints that a solution must satisfy (e.g. on the values of variables) are likely to be easier than those where the task has many constraints that must be satisfied or there are conflicting constraints.
Distance to goal	The greater the number of steps needed to solve the task (reach the goal state), the more difficult the item is likely to be.
Reasoning skills required	The difficulty of an item is influenced by the complexity and type of reasoning skills involved in solving the task.
Restriction on experimentation	It is likely that the less opportunity there is to experiment in a scenario, the harder an item becomes; in other words, unlimited experimentation will make such an item relatively easier.
Unfamiliarity	If a scenario is unfamiliar to a problem solver, the solver may feel ill-equipped to tackle the problem.
Learning	If several items are based around a common scenario, later items are likely to be easier since the solver will have learned something about the system in attempting the earlier items.

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

Osman remarks that "… direct comparison of CDC [Complex Dynamic Control] tasks on multiple dimensions of task complexity has yet to be investigated…" (Osman, 2010). The current study partially addresses this need and includes both the interactive and the static problems used in PISA 2012.

The structure of interactive problems

Fischer et al. (2012) state that in the literature on complex problem solving, "it is mostly the structure of the external problem representation that is considered complex. So a problem usually is considered being of a certain complexity, even if it might seem less complex to problem solvers with more expertise". What factors characterise the structure – and therefore influence the difficulty – of an interactive problem? How do these factors relate to the task characteristics affecting item difficulty proposed above?

Most of the interactive problems in this study were built on mathematical models whose parameters can be varied to realise differing degrees of item difficulty. The two main paradigms used were linear structural equations and finite state machines (FSMs). These have been used extensively in problem-solving research (see Funke, 2001; Buchner and Funke, 1993).

Greiff and Funke (2008) discuss how the formal parameters of MicroDYN systems – systems based on linear structural equations – affect problem difficulty. Examples of these parameters include the number of variables, the quantity and nature of the relationships between them, and the starting and target values for these variables. Greiff et al. (2014) report on a recent study where two of these task characteristics were used to systematically affect item difficulty. In the language of the current study, most of these parameters come under the category of "system complexity".

In a similar manner, one can posit structural attributes potentially affecting difficulty in FSM items. These might include: the number of states in the underlying network, the number of connections between states, the distribution (dispersion) of connections between states, the number of different input signals available, the number of outputs, whether autonomous (time-driven) state transitions can occur without user input, whether "malfunctions" (inconsistencies in behaviour of state transitions) can occur, the relative positions of starting and target states in the network, and how much feedback is given regarding which state the user is in or how close the solver is to the goal. Many (but not all) of these attributes are analogous to the list for MicroDYN systems in Greiff and Funke (2008), and again can be considered aspects of "system complexity".

As part of a detailed model predicting item difficulty in MicroDYN or FSM systems one would need to be able to measure and control the weights of the individual system complexity factors in assessment items. This appears to be more difficult for FSMs than for linear structural equation systems, as the factors are more diverse and some are more difficult to quantify, such as how easy it is to distinguish one state from another based on the output signals (see Neubert et al., 2015 for some recent work in this area). Of course the other factors from Table 9.1 such as how unfamiliar a scenario is, or non-disclosure of information, will affect difficulty in both types of system.

The study described here does not attempt to take all these features into account: indeed, many do not apply to static problems and the item pool was developed to cover both interactive and static problems, capable of being solved by 15-year-olds in a short time. It is acknowledged that more detailed empirical research is warranted on factors affecting item difficulty of linear structural equation systems and especially FSMs, which can be used to model many systems prevalent in everyday life.

Another approach to analysing the structure of complex problems is provided by Quesada et al. (2005). They give a comprehensive taxonomy of CPS tasks as part of an attempt to define CPS. Some of the factors they present are represented in Table 9.1. Quesada et al. list additional factors that include, among many others: event-driven versus clock-driven, stochastic versus deterministic and delayed feedback versus immediate feedback. Again, these generally only apply to interactive problems. It is indeed likely that such factors influence item difficulty in interactive problem situations: their inclusion in further research seems warranted.

Structure of static problems

In general static problems are not based on formal models; no examination of the structure of static problems was considered in this study.

The study

The 10 characteristics from Table 9.1 were posited as factors that influence item difficulty in the 42 problem-solving tasks. Each item was rated by the authors as to the relative "amount" of each of the ten task characteristics[1] it possessed. It should be noted that the rating process was somewhat subjective, and required a good understanding of the underlying problem, how it can be solved and how the characteristics can be identified in the item. In an effort to increase objectivity, each characteristic was described at four levels: 0, 1, 2 and 3. The rating process then consisted of matching each item to the closest description for each characteristic. These levels are defined and discussed in the next section.

The items were administered to students as part of the PISA 2012 assessment in which a total of 85 714 students around 15 years of age were selected from 28 OECD countries and 16 partner economies to obtain a cross-section of each country's student population.[2] All assessment materials were translated into the language of instruction of the students and administered on computer. The students' responses to the items were analysed to obtain estimates of item difficulties.

The study posed two questions.

Research question 1: ***Do the characteristics capture sufficiently different dimensions of cognitive complexity in the items?***

It is certainly likely that there will be dependence between the different characteristics: they will sometimes act in concert and often interact or overlap. Examining the correlations between pairs of characteristics will help to identify whether there is enough difference for them to count as separate dimensions of cognitive complexity. Given the large number of characteristics, it is likely that they will form natural groupings; hierarchical cluster analysis and factor analysis of the characteristics will help identify such groupings.

Research question 2: ***To what extent do the characteristics account for (predict) item difficulty?***

It is expected that these characteristics ought to have a significant impact on how difficult an item is; regression analysis and factor analysis of the item scores associated with their characteristic ratings will assist in quantifying their effects. The outcome of research question 1 will affect the form of the answer: it may turn out to be more useful to consider groups of characteristics (factors) rather than individual ones.

Apart from its intrinsic interest as part of the scientific research into problem solving, there are practical reasons to study the influence of these characteristics. Understanding the difficulty of problem-solving items can assist test developers to:

- develop items with a range of difficulties
- estimate the suitability of test items for test populations
- improve understanding of what is being assessed
- improve judgements about the behaviour of unexpectedly hard or easy items.

Furthermore, understanding the difficulty of problem-solving items can help educators to pinpoint students' strengths and weaknesses, thus helping to identify what further teaching is required, and help explain to students what strategies to use in problem solving.

Described levels for task characteristics

The ten characteristics from Table 9.1 are each described below in terms of four graduated levels. The assumption to be investigated is that higher levels of the characteristic in a task makes the item more difficult.

Amount of information

Level 0: Information content can be understood in one reading. It tends to consist of short, easy-to-read sentences.

Level 1: Some of the information content may require a second reading to understand it. This could be due to the presence of somewhat complex sentences, either long or dense.

Level 2: Some of the information content is difficult to comprehend so several readings through the text may be required. It may consist of complex, dense sentences. Relevant information might need to be selected and linked.

Level 3: There are large amounts of complex information spread throughout the material; this information must be evaluated and integrated in order to understand the problem scenario and task.

Representation of information

Level 0: Solvers must directly use a simple, familiar representation where minimal interpretation is required,.e.g., reading some text or looking up a timetable to see what time the next train leaves.

Level 1: Solvers must interpret a single familiar representation (or an unfamiliar representation that is easily understood) to draw a conclusion about its structure or link two or more familiar representations to achieve a simple goal.

Level 2: Solvers must understand and use an unfamiliar representation that requires substantial decoding and/or interpretation or link an unfamiliar representation to a familiar representation to achieve a goal.

Level 3: Solvers must understand and integrate two or more unfamiliar representations, at least one of which requires substantial decoding and interpretation.

Non-disclosure of information

Level 0: All relevant information required to solve the task is disclosed at the outset.

Level 1: Some simple information (data or rules) necessary to the solution of the task must be discovered by exploring the system. The exploration is scaffolded or fairly random trial and error will suffice to uncover the information.

Level 2: A moderate amount of data and/or rules necessary to the solution of the task must be discovered by exploring the system. The exploration is not directed; a simple but systematic exploration strategy must be formulated and executed.

Level 3: Complex data or rules about the system must be discovered using strategic exploration as part of the solution process.

System complexity

Level 0: The task involves understanding the relationship between a small number of elements/components in the problem situation. The elements are related in a straightforward, one-to-one or linear way.

Level 1: The task requires understanding the relationship between elements/components in the problem situation that are interrelated, possibly in a many-to-one fashion; but it is still possible to treat the relationship between pairs of elements in isolation.

Level 2: The task requires understanding the relationship between elements/components in the problem situation that are interrelated in a many-to-many fashion and/or in an unexpected way (e.g., self-dependencies, or in the terminology of Frensch and Funke, 1995, "eigendynamics"). Some understanding might be required of the way the system works as a whole.

Level 3: The task requires a coherent understanding of all elements or components in the problem situation/system. There are many elements and/or complex interrelations, which may include side effects (a change in one variable causes a change in another).

Constraints to be satisfied

Level 0: There is at most one obvious, explicit constraint that a solution must satisfy.

Level 1: There are two or more explicit, independent constraints that a solution must satisfy or one constraint that requires some interpretation about how to apply it.

Level 2: There is a complex set of interacting/dependent constraints that a solution must satisfy. It might be necessary to set sub-goals that partially satisfy the constraints.

Level 3: There are multiple, conflicting constraints that may be implicit and must be prioritised during the problem solving process.

Distance to goal

Level 0: The goal can be reached in a few steps using a limited number of operators or actions or by repeating a single action a moderate number of times.

Level 1: The goal can be reached in a moderate number of steps by persisting with a set of operators or actions in a linear fashion, or by performing a simple set of actions on different objects/elements or values, then comparing results.

Level 2: The goal can be reached in a moderate number of steps by using a set of operators or actions in a sustained effort where some backtracking is likely.

Level 3: The goal can only be reached in a large number of steps by using a diverse set of operators or actions where backtracking is highly likely, much persistence is required and it may not be clear what the distance is from the current state to the goal state.

Reasoning skills required

Level 0: Only basic reasoning skills are required, such as sorting, classifying or summarising.

Level 1: Inferences need to be made to link related pieces of information; expressions or conditions must be evaluated.

Level 2: Sustained reasoning is required during the course of solving the problem posed by the task or disparate pieces of information must be logically connected. High-level reasoning (e.g. combinatorial reasoning) may be involved.

Level 3: Chains of reasoning must be made to solve the problem posed by the task. Conclusions need to be justified and hypotheses systematically tested. The task may involve more than one type of reasoning.

Restriction on experimentation

Level 0: The scenario allows unlimited experimentation or trying out of state possibilities. It is easy to explore the whole problem space and the effect of actions is clear.

Level 1: The scenario allows experimentation or trying out of state possibilities. It is easy to explore parts of the problem space, but some barriers may exist making it harder to explore the whole space (e.g. some pathways might be blocked or there might be missing information that must be discovered.) The effect of actions is usually clear.

Level 2: The scenario allows some experimentation or trying out of state possibilities, but these are constrained or limited so that it may not be possible to explore all of the problem space.

Level 3: Either the scenario does not allow experimentation, or experimentation is not helpful in working towards a solution.

Unfamiliarity

Level 0: The scenario only involves familiar everyday objects or concepts typically encountered by the solver at home, school or with peers.

Level 1: The scenario involves objects or concepts that might not have been used by the solver before but which are recognisable as belonging to the everyday world.

Level 2: Some key elements in the scenario are unlikely to have been encountered by the solver before and require some thought in order to be understood.

Level 3: The scenario contains unusual or abstract elements that are likely to be difficult for the solver to relate to his or her everyday experience. The significance of these elements may be understood only after the solving of the problem is underway.

Learning

There were either two or three tasks presented in each of the units; and the units were administered in lock-step fashion on the computer, so that the order in which a student attempts a task is fixed. The rating for learning was assigned systematically, decreasing from 3 to 1, for each successive task in a unit (depending on how many tasks there are in the unit).

Sample items with ratings

Ratings for two items from released units that were included in the field trial for the PISA 2012 assessment are given below.

The first item is MP3 Player Question 2, which is discussed in detail in Chapter 5. Briefly, the solver must set a (simulated) MP3 player to a particular state (with individual settings for music type, volume and bass). There are two score points available: full credit is obtained by reaching the goal in (close to) the minimum possible number of steps. Partial credit is obtained by reaching the goal in any number of steps. (A step here is a button click on the MP3 player.)

The second item to be rated is from another unit, Birthday Party, as shown in Figures 9.1 and 9.2.

The scenario for this unit involves guests at a birthday party who must be placed around a dinner table – using "drag and drop" – in a way that satisfies nine specified conditions. This unit is static as all the information necessary to solve the problem is given at the outset. Note that all the *implications* of the conditions are not immediately apparent, however.

It is instructive to compare the authors' consensus ratings for these two items using the task characteristics from Table 9.1, as shown for each task in Table 9.2. Note that in assigning ratings for a characteristic, the range of difficulties present in the entire problem set was kept in mind.

Figure 9.1. **Birthday Party: Stimulus interactive area**

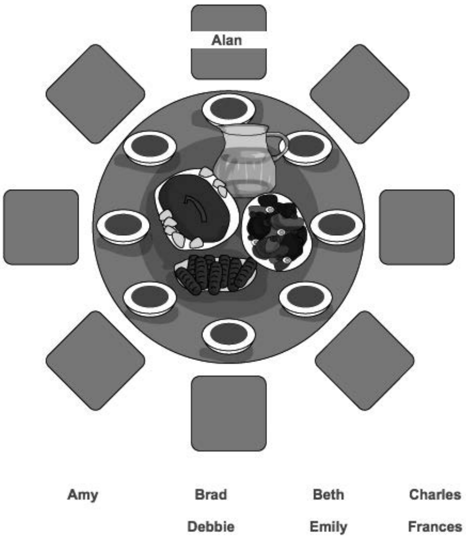
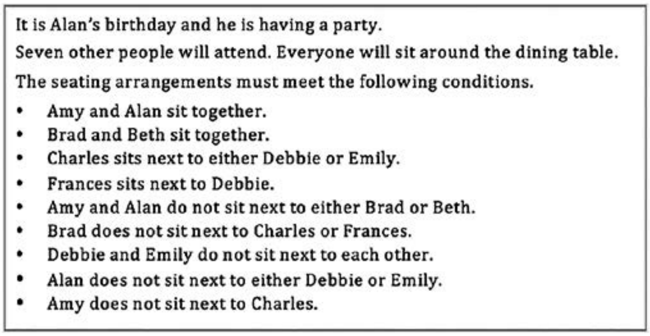


Figure 9.2. **Birthday Party: Information provided**



It is Alan's birthday and he is having a party.

Seven other people will attend. Everyone will sit around the dining table.

The seating arrangements must meet the following conditions.

- Amy and Alan sit together.
- Brad and Beth sit together.
- Charles sits next to either Debbie or Emily.
- Frances sits next to Debbie.
- Amy and Alan do not sit next to either Brad or Beth.
- Brad does not sit next to Charles or Frances.
- Debbie and Emily do not sit next to each other.
- Alan does not sit next to either Debbie or Emily.
- Amy does not sit next to Charles.

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

Table 9.2 **Ratings for MP3 Player Item 2 and Birthday Party Item 1**

Characteristic	**Value (MP3 Player Item 2)**	**Value (Birthday Party Item 1)**
Amount of information	0	0
Representation of information	1	0
Non-disclosure of information	2	0
System complexity	2	0
Constraints to be satisfied	1	2
Distance to goal	2	2
Reasoning skills required	1 (full credit)	1
	0 (partial credit)	
Restriction on experimentation	3 (full credit)	0
	0 (partial credit)	
Unfamiliarity	0	0
Learning	2	3

Since MP3 Player Item 2 is an interactive problem, its rating for non-disclosure of information must be greater than 0. It was in fact given a fairly high rating (2) since no instructions were provided and it takes careful, strategic exploration to work out how to operate the player efficiently. The complexity of the underlying finite state machine is also fairly high (system complexity = 2), presumably serving to further increase the difficulty of this task. The amount of information and unfamiliarity are both 0, as they are not likely to contribute substantially to the item's difficulty. Representation of information and constraints to be satisfied (both 1) are a little more significant, but predicted to have only a moderate effect on the item's difficulty. The reasoning skills required and the restriction on experimentation are scored differently for the two different levels of credit. In order to get full credit, some careful thought (reasoning) is required to ensure that very few mistakes are made in reaching the goal; this also means that additional experimentation won't help (restriction on experimentation = 3). In contrast, enough experimenting will almost certainly lead to the obtaining of partial credit (restriction on experimentation = 0). MP3 Player Item 2 receives 2 for learning since it is the second task in the unit.

Birthday Party Item 1 is a static problem, so its rating for non-disclosure of information must be 0. System complexity is also 0 since there are no underlying states or variables in the problem situation (apart from, perhaps, places around a table). The difficulty of this item lies in the large number of conditions imposed (constraints to be satisfied = 2) and the skills required to monitor and adjust partial solutions relative to these constraints until a complete solution is found (distance to goal = 2, reasoning skills required = 1). Birthday Party Item 1 is made easier by virtue of the fact that solvers can drag and drop peoples' names around the table as much as they like: this experimentation is predicted to assist a solver in working towards a solution. Learning in Birthday Party Item 1 is rated 3, since it is the first item in the unit.

Without considering how the characteristic ratings might be weighted, judging from the generally higher ratings received by MP3 Player Item 2, it may be predicted that it is more difficult than Birthday Party Item 1. The field trial results bear this out, but it was only found to be modestly more difficult, by about 0.3 logits.

Analysis and results

An item response theory (IRT) analysis of student responses was completed for the 42 problem-solving items in the PISA 2012 problem-solving assessment item set. The results showed that there was a good spread of difficulties, from easy items through to hard ones.

Inter-rater reliability analysis

Each of the problem-solving tasks was rated by the authors individually using the four-point scale described above for the amount of each characteristic it possessed, based on how important a part each was believed to play in making the problem task difficult to solve. An inter-rater reliability analysis was initially used to examine the consistency of these ratings. There was found to be a satisfactory level of agreement of ratings across the characteristics (from 0.94 for non-disclosure of information, down to 0.56 for distance to goal, using Kendall's Tau-b statistic which adjusts for tied ratings). Discrepancies were discussed and reconciled to obtain a final consensus rating. This took place over a period of a year: the characteristics' descriptions and levels were refined to make them easier to apply and items were re-rated, in an iterative process.

Regression analysis

The data were analysed by performing a regression analysis on the ten characteristics. This analysis was performed to identify the characteristics having the most effect in predicting the amount of variance in item difficulty. The results are shown in Table 9.3.

Table 9.3. **Standardised regression coefficients**

Model characteristic	Standardised coefficients (Beta)	t	Significance
Amount of information	.205	1.494	.142
Representation of information	.049	.333	.741
Non-disclosure of information	-.187	-1.872	.068
Unfamiliarity	.033	.338	.737
System complexity	.050	.515	.609
Constraints to be satisfied	.188	1.846	.071
Distance to goal	.175	1.443	.156
Reasoning skills required	.500	3.911	.000
Learning	.215	1.989	.053
Restriction on experimentation	.529	5.534	.000

Collectively, these ten variables explained 65.7% (adjusted R^2 value) of the variance in the observed difficulty scores of the items. The related analysis of variance for the model showed that the regression model derived was statistically significant ($p < 0.001$) in predicting more of the variability than expected by chance. It can be seen from Table 9.3 that that the factors having the most effect in the model are reasoning skills required and restriction on experimentation. Note that the table shows coefficients after standardising all variables to have a variance of 1, to allow for different metrics being applied when rating the characteristics. It was observed that this model is somewhat sensitive to changes in rating values. For example, when mistakes in eight learning ratings were corrected (increasing each by 1) and the analysis re-run, the standardised coefficient for reasoning skills required dropped 0.026 and restriction on experimentation increased by 0.006, thereby reversing the order of significance of these two characteristics.

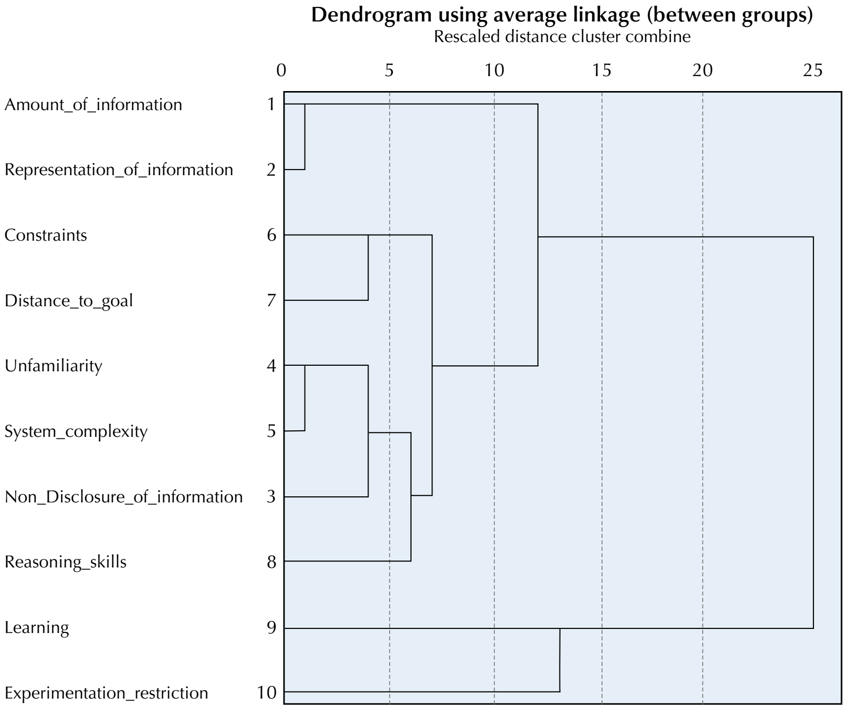
The coefficient for non-disclosure of information seemed counter-intuitive, since it was expected that greater non-disclosure of information would lead to greater difficulty. One explanation for this is that items requiring uncovering of information were scaffolded to varying degrees, to lead test-takers to a response quickly: an average time of around two minutes per item was allowed. Removing non-disclosure of information from the model in addition to some of the other characteristics that were seen to have less significance led to a more parsimonious model with almost as much explanatory power. Using just reasoning skills required, restriction on experimentation and constraints to be satisfied still explained 61.6% of variance. This suggests that there is quite a degree of dependence between the characteristics.

Cluster analysis

To examine the way in which the rating data are related, a hierarchical cluster analysis was conducted. Clustering items by the expert ratings of the ten characteristics produces the tree shown in Figure 9.3. The average linkage routine was used to amalgamate the variables into the dendrogram, based on the average distance from all items in one cluster to another (Hair et al., 1998).

First to link in the hierarchy were amount of information and representation of information, closely followed by unfamiliarity and system complexity. Next were constraints to be satisfied and distance to goal, then non-disclosure of information joined unfamiliarity and system complexity, and so on.

Figure 9.3. **Dendrogram showing clustering of 10 item characteristics**



Overall, this picture of how the characteristics clustered provides information that can help with further parametric-based analyses and in particular suggests groupings of characteristics into a smaller number of factors.

Correlations between variables

The next analysis was the establishment of the Pearson product-moment pair-wise correlations amongst the characteristics in Table 9.1 together with item difficulty. The pairs of characteristics with the highest absolute correlations were:

- amount of information and representation of information (0.759)
- reasoning skills required and difficulty (0.618)
- reasoning skills and distance to goal (0.541)
- restriction on experimentation and difficulty (0.525)
- reasoning skills required and learning (-0.493)

Also of interest was that distance to goal had close to a zero correlation with amount of information, representation of information, non-disclosure of information and unfamiliarity.

This sheds some light on research question 1: do each of the characteristics capture sufficiently different dimensions of cognitive complexity in the items? It can be seen that there is a good deal of dependence between the characteristics. However, whilst there are significant correlations at the 0.001 level, the highest value of any correlation is 0.759 followed by 0.618. So the correlations are quite a distance from being 1. Taken with the fact that some characteristics had correlations approaching 0, it can be seen that the characteristics do cover a range of dimensions.

High correlations should not be taken to mean that some characteristics are redundant, for two reasons. First, some greater correlations might have resulted because of the way the items were designed in this particular PISA assessment problem set. It seems reasonable to believe that, if desired, a test developer could reduce the correlation between any pair of characteristics by deliberately rewriting items to do so. Second, it may be that greater correlations are a result of the rating process, wherein the rater has perhaps unconsciously associated two characteristics, for example, number of representations and amount of information. This cause of correlation can be minimised by tighter definitions of the characteristics and more care in assigning the ratings.

Factor analysis

The final analysis was a factor analysis using the ten variables from Table 9.1 as predictors of the difficulty values for the items.

A principal component factor analysis of the correlation matrix was performed, followed by a varimax rotation with Kaiser normalisation to determine the maximal separation of the defining factor loadings (see Kaiser, 1958). Four factors with eigenvalues greater than one emerged which together explained a total of 72% of the variance. The loadings for these four factors are shown in Table 9.4.

Table 9.4. **Rotated component matrix**

Characteristic	**Factor**			
	1	2	3	4
Amount of information	0.00	**0.92**	–0.05	–0.01
Representation of information	0.08	**0.90**	–0.10	0.14
Non-disclosure of information	0.07	0.52	0.54	0.24
Unfamiliarity	–0.11	0.17	–0.03	**0.87**
System complexity	0.43	–0.02	0.17	0.66
Constraints to be satisfied	0.30	0.30	**–0.70**	0.14
Distance to goal	**0.81**	–0.12	–0.28	0.17
Reasoning skills	**0.77**	0.27	0.29	0.25
Learning	**–0.82**	–0.05	0.06	0.16
Restriction on experimentation	0.06	0.00	**0.68**	0.09

It is usual practice with confirmatory factor analysis to require loadings to be of size 0.7 or more to confirm that the variables are represented by a particular factor. By this criterion, most characteristics fitted into a factor, the clear exception being non-disclosure of information. The factors consist of the following groups of task characteristics (highlighted in Table 9.4), in order of decreasing explanatory power:

1. **Factor 1** (27.9%): learning, distance to goal, reasoning skills required
2. **Factor 2** (18.8%): amount of information, representation of information
3. **Factor 3** (15.1%): constraints to be satisfied, restriction on experimentation
4. **Factor 4** (10.2%): system complexity, unfamiliarity

This result is supported by the hierarchical cluster analysis and the Pearson correlations given above, which gave similar groupings and addresses research question 2 in showing the relative extent to which the characteristics account for item difficulty.

Discussion

It is fruitful to consider *why* the characteristics are grouped into these four factors. Are there organising principles at play here? Looking at the constituent characteristics of each factor, some tentative classifications (interpretations) can indeed be made. Note that the following discussion goes beyond the data and involves an interpretation of the meaning of the characteristics. This should be worthwhile, because in addition to helping explain the existence of the factor groupings, the classifications can be used as an aid in test construction.

Factor 1 – Barriers

Solving a problem includes overcoming or avoiding barriers between a given state and a goal state by applying an appropriate transformation. The exercise of particular types of reasoning is required to perform such transformations, for example, using combinatorial reasoning to generate possible paths and select an optimal one, or logical reasoning to deduce relevant conclusions from given information. The sheer number of transformations required can affect the amount of reasoning required, so distance to goal combines with reasoning skills required in this factor.

Learning is part of Factor 1, but its loading is negative: higher values for reasoning skills required go with lower values for learning. This may simply be a test artefact: the items in a problem unit requiring more reasoning tend to be later in the unit (and thus have a lower value for learning).

Factor 2 – Information

The characteristics in this factor are clearly about information, both the amount of information that the problem solver must process to understand the problem and the form of the information – how it is represented – which might include text, graphs, images or tables.

Factor 3 – Constraints

High values for constraints to be satisfied tend to go with lower values for restriction on experimentation. Recall that a low value on restriction on experimentation our prediction was that the ability to experiment was likely to have made the item easier. The fact that these variables have loaded on the same factor, but are opposite in sign might indicate that one way to deal with difficult constraints in a problem is to try out possibilities if this is allowed. Birthday Party Item 1 illustrates this point (see earlier discussion and Table 9.2) and strongly suggests that if studies such as Turner (2012) are extended to computerised assessments then the amount of experimentation allowed may need to be taken into account.

Factor 4 – System features

To solve a problem a person must build a mental model of the system features, including its underlying structure. Complex systems tend to be unfamiliar, which may explain why system complexity and unfamiliarity have loaded on the same factor.

Possible improvements and future directions

A potential criticism of the study is that the ratings for the amount of a characteristic present in an item are somewhat subjective, thereby affecting the reliability of the results. Although training and experience, for example, will make a difference in the accuracy and reliability of a rater's results, it would be more satisfying if there were more objective measures of system complexity, distance to goal etc. This might be possible for systems based on mathematical models: for some work in this direction with MicroDYN systems, see Greiff and Funke (2008), Kluge (2008) and Greiff (2012). Greiff found that the two factors with the most influence on MicroDYN item difficulty were the number of effects between the variables and the quality of effects (i.e. whether there are side effects or "autoregressive processes"). Kluge's work supports this finding. These "effect" characteristics fall under the umbrella of Factor 4: system features in the present study.

It might be fruitful to carry out an analysis similar to the current study for a set of problems based solely on finite state machines, using a set of characteristics such as those suggested in the section above on the structure of interactive problems. This is especially interesting because of the prevalence in everyday life of systems based around FSMs, and would enable better face validity of problem scenarios.

The regression model was able to account for 65.7% of the variance in item difficulty and the factor analysis explained 72% of the variance. Further refinements of the characteristics based on the above considerations of a system's structure should serve to improve the predictive accuracy of the model. Even if the measurement errors in modelling item difficulty using IRT are not too large, there still remains the possibility of the effect of individual differences among problem solvers: different strategies used by different individuals might tap into different characteristics. This would make the ratings dependent on the population, not just the items.

Conclusion

Despite some shortcomings of the ratings system and the somewhat ad hoc interpretation of the factors, some useful results emerge. At a basic level, it can be seen for example that changing barrier characteristics (by increasing the number of steps required to reach a goal, for instance) will tend to increase the difficulty of a problem. At a more general level, the analysis has shown that collectively the four factors identified account for over 70% of the variability in item difficulties, given the ten characteristics. Further, the characteristics that individually best model the difficulty of items are reasoning skills required, restriction on experimentation and constraints to be satisfied, which account for 61.6% of the variance.

Test developers could systematically vary the four factors (barriers, information, constraints and system structure) across problem-solving assessment items, so as to facilitate the construction of assessments that span a range of difficulty levels, or that target particular populations.

Notes

[1] "Task characteristics" will be simply referred to as "characteristics".

[2] For the PISA 2012 main survey a scientific two-stage sampling method was used (see OECD, 2012). For the present study the abilities of the students sitting the assessment is relatively unimportant.

References

Buchner, A. and J. Funke (1993), "Finite-state automata: Dynamic task environments in problem-solving research", *Quarterly Journal of Experimental Psychology*, Vol. 46A/1, pp. 83-118, http://dx.doi.org/10.1080/14640749308401068.

Ewert, P.H. and J.F. Lambert (1932), "Part II: The effect of verbal instructions upon the formation of a concept", *Journal of General Psychology*, Vol. 6/2, pp. 400-411.

Fischer, A., S. Greiff and J. Funke (2012), "The process of solving complex problems", *The Journal of Problem Solving*, Vol. 4/1, pp. 19-41, http://docs.lib.purdue.edu/jps/vol4/iss1/3.

Frensch, P.A. and J. Funke (eds.) (1995), *Complex Problem Solving: The European Perspective*, Erlbaum, Hillsdale, NJ.

Funke, J. (2001), "Dynamic systems as tools for analysing human judgement", *Thinking and Reasoning*, Vol. 7/1, pp. 69-89, http://dx.doi.org/10.1080/13546780042000046.

Funke, J. and P.A. Frensch (1995), "Complex problem solving research in North America and Europe: An integrative review", *Foreign Psychology*, Vol. 5, pp. 42-47.

Greiff, S. (2012), *Individualdiagnostik komplexer Problemlösefähigkeit* [Assessment of Complex Problem-Solving Ability], Waxmann, Munster.

Greiff, S. and J. Funke (2008), "What makes a problem complex? Factors determining difficulty in dynamic situations and implications for diagnosing complex problem solving competence", in J. Zumbach, N. Schwartz, T. Seufert and L. Kester, (eds.), *Beyond knowledge: The legacy of competence. Meaningful computer-based learning environments*, Springer Science, Dordrecht, pp. 199-200, http://-dx.doi.org/10.1007/978-1-4020-8827-8_27.

Greiff, S., K. Krkovic and G. Nagy (2014), "The systematic variation of task characteristics facilitates the understanding of task difficulty: A cognitive diagnostic modeling approach to complex problem solving", *Psychological Test and Assessment Modeling* Vol. 56/1, pp. 83-103.

Hair, J.F., R.E. Anderson, R.L. Tatham and W.C. Black (1998), *Multivariate Data Analysis*, 5th Edition, Prentice Hall, Upper Saddle River, NJ.

Kaiser, H.F. (1958), "The varimax criterion for analytic rotation in factor analysis", *Psychometrika*, Vol. 23/3, pp. 187-200.

Kluge, A. (2008), "Performance assessments with microworlds and their difficulty", *Applied Psychological Measurement*, Vol. 32/2, pp. 156-180, http://dx.doi.org/10.1177/0146621607300015.

Lumley, T., A. Routitsky, J. Mendelovits and D. Ramalingam (2012), "A framework for predicting item difficulty in reading tests", paper presented at the Annual Meeting of the American Educational Research Association (AERA), Vancouver, British Columbia, Canada, 13-17 April 2012, http://works.bepress.com/dara_ramalingam/2.

Masters, G.N. and B.D. Wright (1997), "The partial credit model", in W.J. van der Linden and R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory,* Springer, Heidelberg, pp. 101-122.

Mayer, R.E. (1998), "Cognitive, metacognitive, and motivational aspects of problem solving", *Instructional Science*, Vol. 26/1-2, pp. 49-63.

Mayer, R.E. (1992), *Thinking, Problem Solving, Cognition*, 2nd edition, Freeman, New York.

Neubert, J.C., A. Kretzschmar, S. Wüstenberg and S. Greiff (2015), “Extending the assessment of complex problem solving to finite state automata: Embracing heterogeneity”, *European Journal of Psychological Assessment*, Vol. 31/3, pp 181-194.

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

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

OECD (2012), *PISA 2009 Technical Report*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264167872-en.

Osman M. (2010). “Controlling uncertainty: A review of human behavior in complex dynamic environments”, *Psychological Bulletin,* Vol. 136/1, pp. 65-86.

Quesada, J.F., W. Kintsch and E. Gomez (2005), “Complex problem solving: A field in search of a definition?”, *Theoretical Issues in Ergonomic Science*, Vol. 6/1, pp. 5-33.

Rasch, G. (1960), *Probabilistic Models for some Intelligence and Attainment Tests*, Nielsen and Lydiche, Copenhagen

Turner, R. (2012), “Some drivers of test item difficulty in mathematics”, paper presented at the Annual Meeting of the American Educational Research Association (AERA), Vancouver, British Columbia, Canada, 13-17 April 2012, http://works.bepress.com/ross_turner/17.

Turner, R., et al. (2013), “Using mathematical competencies to predict item difficulty in PISA: A MEG study”, in M. Prenzel, M. Kobarg, K. Schöps and S. Rönnebeck, (eds.), *Research on PISA: Research Outcomes of the PISA Research Conference 2009*, Springer Science, Dordrecht, pp. 23-37.

Chapter 10

Assessing complex problem solving in the classroom: Meeting challenges and opportunities

By
Philipp Sonnleitner, Ulrich Keller, Romain Martin
Luxembourg Centre of Educational Testing, University of Luxembourg, Luxembourg.
Thibaud Latour
Luxembourg Institute of Science and Technology, Luxembourg.
and Martin Brunner
Institut für Schulqualität, FU Berlin, Germany.

Complex problem solving (CPS) is now established as a key aspect of today's educational curricula and a central competence for international assessment frameworks. It has become clear that the educational context – particularly among general school-age students – places particular demands on the instruments used to assess CPS, such as computer-based microworlds. This chapter shows how these challenges can successfully be addressed by reviewing recent advancements in the field of complex problem solving. Using the example of the Genetics Lab, a newly developed and psychometrically sound microworld which emphasises usability and acceptance amongst students, this chapter discusses the challenges and opportunities of assessing complex problem solving in the classroom.

This work was supported by funding from the National Research Fund Luxembourg (FNR/C08/LM/06). The authors would like to thank all the students and teachers participating in our studies.

Introduction

It seems beyond doubt that in a world facing challenges like globalisation, global warming, the financial crisis or diminishing resources, the problems our society has to solve will become more complex and difficult during the next years. In their mission to prepare younger generations to successfully respond to these enormous challenges, schools have to adapt too. Therefore, it is not surprising that many contemporary educational curricula and assessment frameworks like the Programme for International Student Assessment (PISA; OECD, 2004, 2013) stress the integration and assessment of the ability to solve (domain-general) complex and dynamic problems (Leutner et al. 2012; Wirth and Klieme, 2003). In order to achieve this, many scholars suggest the use of computer-based problem-solving scenarios, so-called “microworlds” that allow the tracking of students’ problem-solving process as well as their problem representations (Bennett, Jenkins, Persky and Weiss, 2003; Ridgway and McCusker, 2003) – crucial information for interventions aimed at increasing problem-solving capacity in students.

Surprisingly, despite great enthusiasm for microworlds in the educational field, most previous studies have drawn on adults, typically of high cognitive capacity (e.g. university students of various kinds). Only a few studies so far have directly used such microworlds and investigated their psychometric properties among populations of school students. These exceptions, however, mainly focused on students of the higher academic track, and usually at 10th grade or above (e.g. Kröner, Plass and Leutner, 2005; Rigas, Carling and Brehmer, 2002; Rollett, 2008; Süß, 1996). Due to the highly selective samples used by these studies, it is questionable to what extent microworlds can unconditionally be applied to the whole student population without modifications to their construction rationale or scoring procedures.

This chapter identifies and discusses the challenges that arise when microworlds are administered “in the classroom”: the special characteristics of today’s students, also described as “digital natives”, and the need for timely behaviour-based scoring procedures that are at the same time easy to understand by educators and teachers. By taking the Genetics Lab, a microworld especially targeted at students aged 15 and above in all academic tracks as an example (Sonnleitner et al., 2012a), this chapter presents opportunities to react on these challenges, based on three independent studies using the Genetics Lab.

The Genetics Lab: A microworld especially (PS) developed for the classroom

To learn more about the application of microworlds in the educational setting and to further investigate to what extent and in what way microworlds have to be adapted for this context, the Genetics Lab (GL) was developed at the University of Luxembourg (see Sonnleitner et al., 2012a, 2012b). The goal was to set up a (face-)valid, psychometrically sound microworld to assess complex problem solving (CPS) that can immediately be applied in the school context. To this end, the development drew on the rich body of empirical knowledge derived from previous studies on microworlds (for an overview see, for example, Blech and Funke, 2005; Funke and Frensch, 2007). To enable educators to make full use of the GL, it can be administered within 50 minutes (i.e. the length of a typical school lesson), and in four different languages (English, French, German and Italian). Moreover, it was published under open-source licence and can be freely downloaded and used.[1]

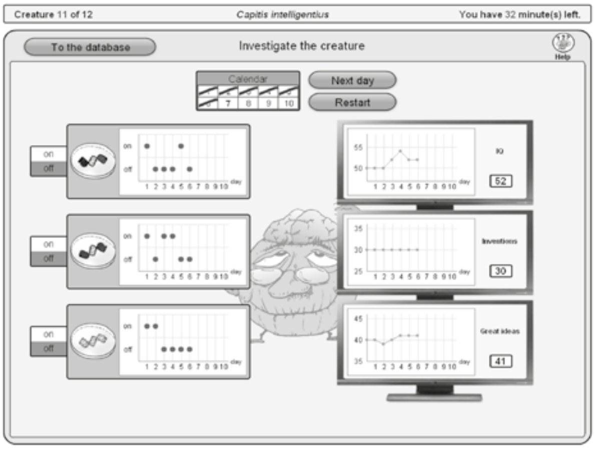
In the GL (shown in Figures 10.1 to 10.3 and described below), students explore how the genes of fictitious creatures influence their characteristics. To this end, students can actively manipulate the creatures’ genes by switching them “on” or “off” and then study the effects of these manipulations on certain characteristics of the creatures (Figure. 10.1). Genes (i.e. input variables) are linked to the characteristics (i.e. output variables) by linear equations. It is the task of the student to find out about these (non-transparent) relations and to document the gathered knowledge (Figure 10.2). Finally, the students have to apply the knowledge they have gathered to achieve certain target values of the creatures’ characteristics (Figure 10.3). These task characteristics allow students to be scored for 1) their exploration and information-gathering behaviour; 2) their gathered knowledge in the form of a causal diagram showing the relations they have discovered between genes and characteristics; and 3) their ability to apply the knowledge to achieve certain target values for the creatures' characteristics. Each creature is designed in a way to realise key features of a complex problem (see for example Funke, 2001, 2003): 1) complexity, by including a high number ofvariables (several genes and characteristics); 2) connectivity, by linking the variables via linear equations; 3) dynamics, by implementing an automatic change of certain characteristics that is independent from the students' actions; 4) intransparency, by hiding the connections between the variables; and 5) multiple goals, by asking the student to achieve different target values on several of the creature's characteristics. For further details about the GL's scores and construction rationale please see Sonnleitner et al. (2012a).

The GL has been applied in 3 independent studies so far with more than 600 participating students (see Table 10.1 for an overview). To foster commitment and motivation, detailed written feedback on students' performance was offered. Further details concerning the first two studies are given in Sonnleitner et al. (2012a), and the third in Sonnleitner, Keller, Martin, & Brunner (2013). The data gathered from these studies, along with the experiences gained, inform the following discussion of the challenges and opportunities of microworlds within the educational field.

Table 10.1. **Sample characteristics of the Genetics Lab studies**

		Study 1	**Study 2**	**Study 3**
n		43	61	563
Mean age (SD)		15.8 (.87)	15.5 (.61)	16.4 (1.16)
Male		24	26	279
Female		19	35	284
School track	intermediate	43	35	234
	academic	–	26	329
School grade	9th	43	61	300
	11th	–		263

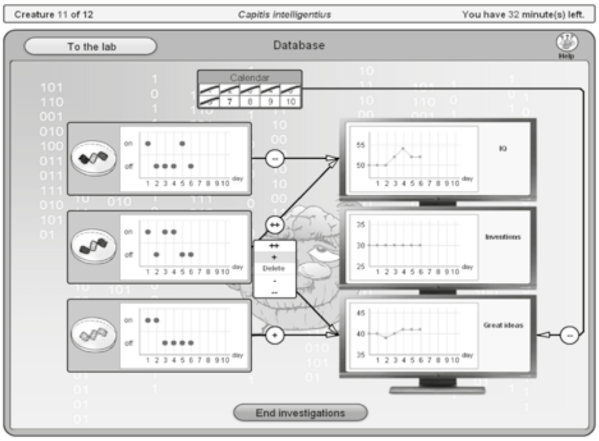
Figure 10.1. **Genetics Lab Task 1: Exploring the creature**



Source: Sonnleitner et al. (2012a), "The Genetics Lab: Acceptance and psychometric characteristics of a computer-based microworld assessing complex problem solving".

Task 1: Exploring the creature. First, students investigate the effects of certain genes on a creature's characteristics. Genes, and thus their effects, can be switched "on" or "off". By clicking the "next day" button at the top of the screen students can progress in time and then observe the effects of their manipulations by studying the related diagrams (genes are depicted in red, characteristics are depicted in green diagrams).

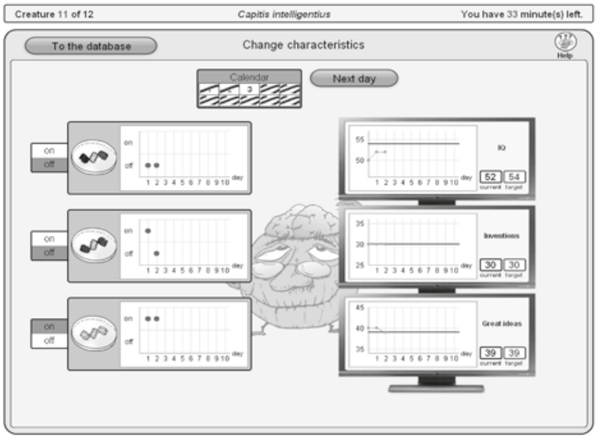
Figure 10.2. **Genetics Lab Task 2: Documenting the knowledge**



Source: Sonnleitner et al. (2012a), "The Genetics Lab: Acceptance and psychometric characteristics of a computer-based microworld assessing complex problem solving".

Task 2: Documenting the knowledge. While exploring the creature, students document the knowledge they have gathered about the genes' effects in a related database that shows the same genes and characteristics as the lab. They do this by drawing a causal diagram indicating the strength and direction of the discovered effects. The resulting model can be interpreted as the students' mental model of the causal relations.

Figure 10.3. **Genetics Lab Task 3: Changing the characteristics**



Source: Sonnleitner et al. (2012a), "The Genetics Lab: Acceptance and psychometric characteristics of a computer-based microworld assessing complex problem solving".

Task 3: Achieving target values. Finally, students have to apply the gathered knowledge in order to achieve certain target values on the creature's characteristics. Crucially, they only have three "days" or time steps left to do this. Students therefore have to consider the dynamics of the problem and plan their actions in advance.

Challenge 1 – Digital natives

A crucial aspect of an assessment instrument is its suitability for the characteristics and background of the target population (AERA, APA and NCME, 1999:131). Previous studies using microworlds mostly sampled adult populations, typically university students. Thus, the question arises whether and in what way today's students differ from these groups.

Today's students were born in the late 1990s and are described as members of the "net generation" (Tapscott, 1998) or are even called "digital natives" (Prensky, 2001a, 2001b), mainly because they have grown up in a world in which information and communication technology (ICT) is permanently available. In contrast with former generations, they have to deal with digital media like video games, simulations, the Internet, instant messaging, virtual learning environments and social networks almost on a daily basis. Hence, some authors have claimed that this interaction with digital media right from birth caused today's students to be cognitively different from prior generations. According to Prensky (2001a, 2001b), digital natives think and process information fundamentally differently. They are used to processing information very fast, apply multi-tasking to achieve their goals, strongly rely on graphics and symbols to navigate and, due to their exposure to video games, they are used to getting instant gratification and frequent rewards. In a review of the information-seeking habits of this generation, Weiler (2005) describes these students as primarily visual learners who prefer to actively engage in hands-on activities instead of passive learning. Veen and Vrakking (2006) highlight the iconic skills (use of symbols, icons and colour coding to navigate within digital environments) that this generation has developed in order to deal with a massive and permanent information overload. Moreover, they perceive technology in a new way, as being merely a tool for various purposes and that has to work flawlessly.

Recent reviews, however, showed that differences between today's students and former generations may be overstated (e.g. Bennett and Maton, 2010). Indeed, several studies showed that this generation is a very heterogeneous sample with varying degrees of digital competence (Li and Ranieri, 2010) and technology use (Margaryan, Littlejohn and Vojt, 2011). Nevertheless, the same studies report that almost every member of this generation uses a mobile phone, a personal computer or laptop and has access to the Internet. Thus, while claims concerning the cognitive uniqueness and homogeneity of this generation may be exaggerated, virtually nobody questions their heavy exposure to and use of digital media and devices.

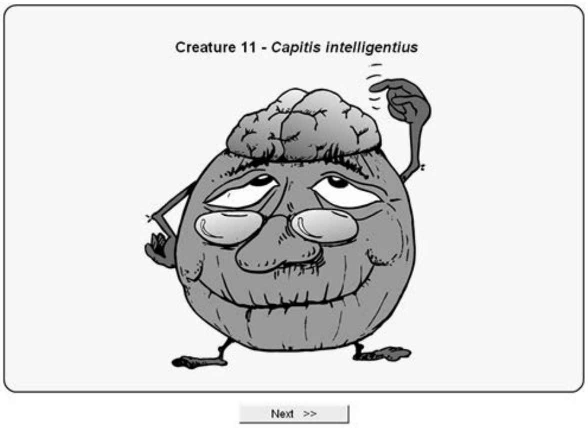
This, in turn, has several crucial implications concerning the expectations of today's students with regard to a computer-based test: First, due to their massive exposure to high-quality (commercial) computer programmes, they expect a perfect and flawlessly functioning technology. Second, especially on the basis of their experiences with video games and newer mobile devices, they expect a completely intuitive graphical user interface (GUI). Students do not want to invest time and effort into working out how to interact with a programme. Third, this GUI should also be appealing and resemble modern standards of design to ensure that the test is perceived as being attractive and of high quality. Fourth, students want to learn how to deal with the task by actively exploring and interacting with it. In contrast, they are very likely to skip any extensive written instructions. Finally, their motivation to interact with a test will be increased by instant gratification or at least instant feedback on their performance. If these criteria are not met by a computer-based test, acceptance of the instrument might be at stake.

Responding to digital natives' needs: Game-like characteristics and usability studies

A first step towards responding to these characteristics of today's students concerned the theoretical conceptualisation of the GL. To start with, we ensured a clear and intuitive GUI by following recommendations on user interface design (e.g. Fulcher, 2003). As can be seen in Figures 10.1 to 10.3, the structure of the GUI clearly resembles the navigation within the GL: the layer at the top shows the progress within the test itself by indicating the number of creatures (i.e. items) that are left and the remaining time available to investigate them. The next layer corresponds to navigating within each item; it provides the buttons to switch between the lab and the database and contains the help function. Finally, all the elements used to directly manipulate the creature or depict the knowledge gathered about it are arranged within the inner layer of the GUI. In addition, elements belonging together (e.g. the calendar and the buttons to progress in time) share the same colour. To make the design of the GL even more appealing and to increase the motivation to work on the test, we implemented several game-like characteristics (see McPherson and Burns, 2007; Washburn, 2003; Wood et al., 2004): A "cover story" was created, putting the student into the role of a young scientist who has started work in a fictitious genetics lab. An older scientist charges the student with investigating several newly discovered creatures and explains the functioning of the lab. Throughout the test, this "virtual mentor" remains present in the form of an integrated help function. In addition, the fictitious creatures are depicted in a funny cartoon-like style and carry humorous characteristics (Figure 10.1 and 10.4). After exploring and manipulating the creature, students get feedback on their performance in the form of two simple scales scoring the causal diagram they created in Task 2 and their control performance in Task 3 (Figure 10.5).

As a reaction to the digital natives' style of learning, we also placed special emphasis on the instructions given at the beginning of the GL. Whereas former microworlds mostly included extensive written instructions or training periods with varying levels of standardisation (Rollett, 2008), the GL's instructions are highly standardised, interactive and based on standards for modern multimedia learning (Mayer, 2003). After a short explanatory text, each task in the GL (exploring the creature, drawing a causal model and achieving target values) is visualised by an animation and has to be practised in a related exercise. As mentioned above, detailed feedback is provided about students' performance in drawing the causal model and achieving the target values.

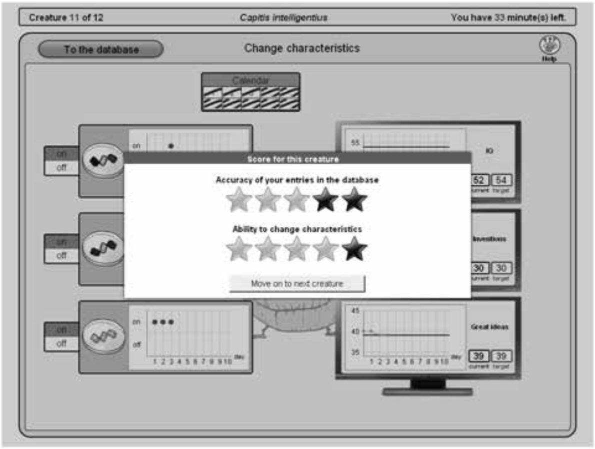
Figure 10.4. **Start screen for a creature**



Source: Sonnleitner et al. (2012a), "The Genetics Lab: Acceptance and psychometric characteristics of a computer-based microworld assessing complex problem solving".

The second step towards ensuring acceptance of the GL among digital natives was to guarantee flawless functioning and high usability. To this end, we adapted and substantially extended traditional test development procedures by including several small-scale usability studies (Figure 10.6). This approach not only aimed at evaluating the design of the GL's GUI in terms of acceptance but also at reducing construct-irrelevant variance in the GL's performance scores (Fulcher, 2003). The participants in the first and second usability studies were both laypeople and experts in the field of testing and usability. The third usability study used university students and students from the target population. All participants were asked to think aloud while working on the GL. Together with these comments, trained observers documented the participants' behaviour, followed by an interview asking participants for any perceived problems and possible solutions. On basis of these data, comprehensibility and functionality problems were identified and discussed in a focus group which prepared suggestions for the modification of the GL. The problems identified ranged from minor problems like a suboptimal position of a button to construct-related problems. For example, it turned out that using the causal diagram for knowledge representation was highly demanding and unfamiliar to 15-year-old students.

Figure 10.5. **Feedback on performance at the end of an item**



Creature 11 of 12 — *Capitis intelligentius* — You have 33 minute(s) left.

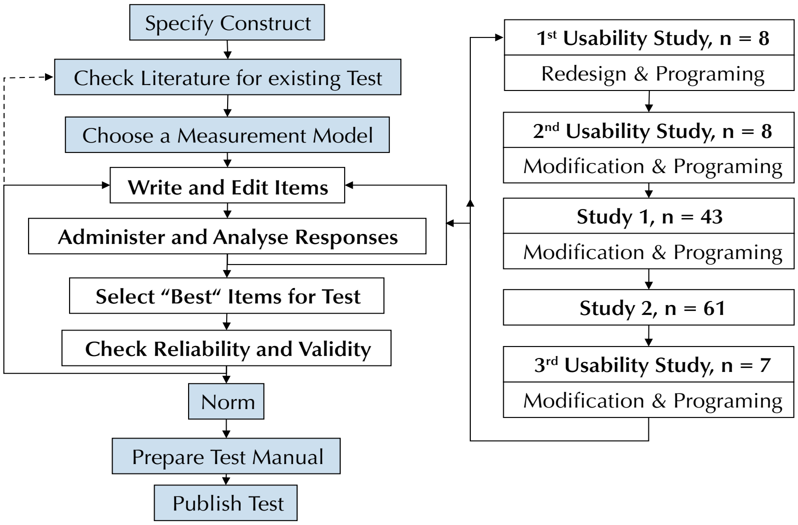
Score for this creature

Accuracy of your entries in the database

Ability to change characteristics

Whereas the results of the first two usability studies caused major revisions in the GUI, especially modifying the wording and sequence of the instructions, the third usability study merely led to minor changes. Importantly, this approach not only warranted high usability of the GL but also led to substantial insights concerning the measured construct and how to derive valid scoring algorithms of students' problem-solving behaviour.

Figure 10.6. **Adapted test development process of the Genetics Lab**



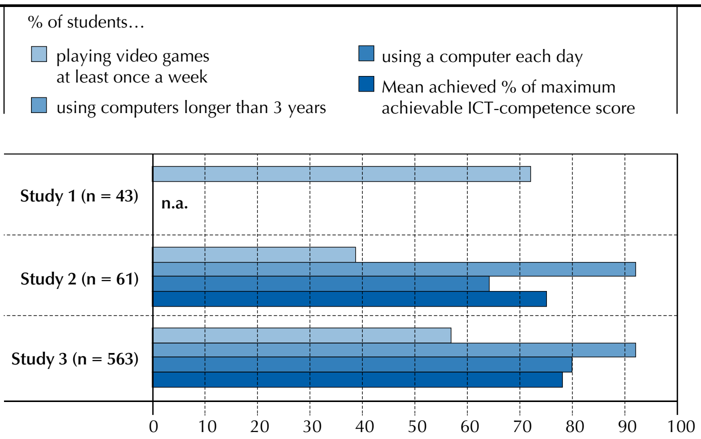
Specify Construct
Check Literature for existing Test
Choose a Measurement Model
Write and Edit Items
Administer and Analyse Responses
Select "Best" Items for Test
Check Reliability and Validity
Norm
Prepare Test Manual
Publish Test
1st Usability Study, n = 8
Redesign & Programing
2nd Usability Study, n = 8
Modification & Programing
Study 1, n = 43
Modification & Programing
Study 2, n = 61
3rd Usability Study, n = 7
Modification & Programing

Source: based on the traditional approach presented by Shum, O'Gorman and Myors (2006), *Psychological Testing and Assessment*.

Acceptance and usability of the Genetics Lab among digital natives

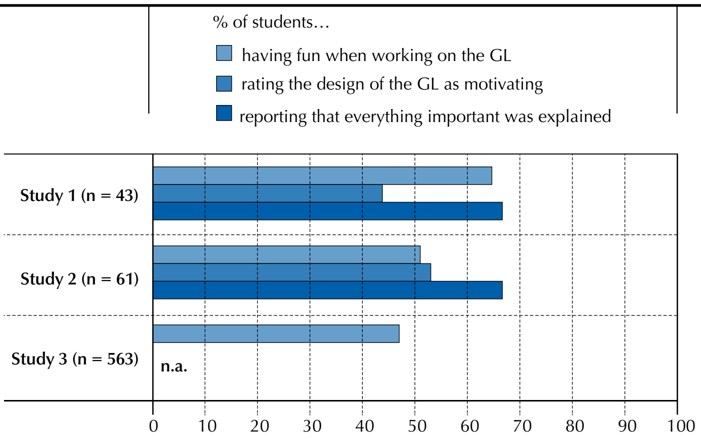
A first analysis of our samples' characteristics showed that the claims made by the literature about today's students were well supported by our studies. Figure 10.7 presents several ICT-related characteristics of the participating students. As can be seen, the vast majority of students had already been using personal computers for more than three years (92%) and nearly every day (up to 80%). Moreover, these students report high levels of ICT-competence on a 10 item questionnaire including several ICT-related activities like burning CDs, downloading pictures and programs from the Internet, creating a webpage or a multimedia presentation. Their total scores on this scale were (linearly) transformed into a percentage of the maximum possible score (POMP) that could be attained (see Cohen et al., 1999). Thus, the percentages depicted in the black bars of Figure 10.7 (75% for Study 1 and 78% for Study 2) describe the mean achieved percentages of a maximum achievable ICT-competence score.

Figure 10.7. **ICT familiarity and competence of students**



Note: n.a. = not applicable.

Figure 10.8. **Acceptance of the Genetics Lab among students**



Note: n.a. = not applicable.

Although results concerning the frequency of video game playing per week are somewhat mixed, looking at the most representative Study 3 indicates that about 57% of the participating students play video games at least once a week. Thus, despite a minority who doesn't use computers on a regular basis and rate themselves as less ICT competent, the vast majority of students can be described as ICT-literate with an extensive experience in dealing with digital environments and computer programs.

To investigate whether our attempt to develop a microworld suited for today's students was successful, we evaluated acceptance of the GL within the conceptual framework of well-established technology acceptance models (e.g. Terzis and Economides, 2011). According to this framework, acceptance of a computer-based assessment instrument may be substantially influenced by its perceived ease of use and perceived playfulness. Moreover, the instrument's comprehensibility and functionality are considered crucial factors determining its usability and hence its acceptance among the target population. Consequently, students participating in Study 1 were asked to rate various elements of the GL in terms of these four dimensions. The results of this questionnaire are presented as POMP scores in Table 10.2. Given the lack of comparable studies, we considered values above 50% – indicating that positive student evaluations outweigh negative evaluations – as positive outcomes (for more details about the questionnaire, please refer to Sonnleitner et al., 2012b). Overall, results show that the GL is well accepted among today's students. The GL was rated as being easy to use (mean = 54, standard deviation = 23) with an attractive appearance (M = 64, SD = 22). Students also attested the GL to be comprehensible (M = 61, SD = 17) and well functioning (M = 60, SD = 22). Moreover, in all three studies, large portions of the students reported having fun while working on the GL (see Figure 10.8). Apart from Study 1, in which the GL was administered on its own, students had to work on the GL at the end of a 2-hour test session. This may explain the somewhat smaller portion of students in Studies 2 and 3 indicating they had fun while working on the GL. Moreover, results show that the game-like design of the GL was appreciated and even described as motivating by large portions of the students. Crucially, the vast majority of students felt that everything important was explained during instructions. We take this as clear indication of the instruction's efficiency at successfully illustrating the handling of the GL in an interactive multimedia-based way.

To sum up, the results suggest that our attempt to develop a microworld with high usability that enjoys high acceptance among today's students was successful. For the first time, we could also go beyond anecdotal evidence that interacting with such scenarios makes fun (e.g. Ridgway and McCusker, 2003). We largely attribute this positive outcome to the actions we have taken to consider special characteristics of today's students, namely, the integration of game-like characteristics, the development of a standardised and interactive multimedia instruction, and extensive usability testing.

Challenge 2 – Scoring

The major reasons for using microworlds in the educational context are 1) assessing and evaluating students' initial CPS skills and to study their relation to other constructs such as general school achievement (Leutner et al., 2012; Wirth and Klieme, 2003); and 2) directly implementing them into educational practice to use them for interventions aiming at improving these skills (Bennett et al., 2003; Ridgway and McCusker, 2003). This in turn poses special challenges concerning the scoring of students' performance on these problem-solving scenarios.

First and foremost, the scores yielded have to be psychometrically sound to allow for reliable and valid score interpretations. Second, in order to be useful for behaviour-based interventions, scores must make full use of the possibilities computer-based testing has to offer. Compared to traditional assessments, which are mostly paper-and-pencil based multiple-choice tests, microworlds allow the capture of digital "traces" left by the student when interacting with these scenarios (i.e. each action of the student is stored in a related log file). Although such traces provide highly valuable information about students' problem-solving behaviour, the scoring of such complex behavioural data is challenging (Hadwin et al., 2007; Winne, 2010). Third, if microworlds are directly used in educational practice to foster CPS skills, it is essential that the interpretation of the performance scores yielded is easy and comprehensive. Educators working with these scores should be able to easily understand and use them to draw sound conclusions about their students' behaviour when confronted with complex problems.

In order to guarantee highly reliable performance scores for the GL, we based its development on the so-called MicroDYN approach (Greiff, Wüstenberg and Funke, 2012). In contrast to former microworlds that consisted of one very extensive problem-solving scenario, each student has to work on several, independent scenarios (i.e. creatures) of varying complexity and content. Thus, when students' performance is aggregated across creatures, the resulting scores of the students' CPS skills are more reliable than those derived from a single scenario. The following sections discuss the GL's performance scores concerning their reliability, internal validity and how the students' traces were used to fully mine the potential of behaviour-based data and to make score interpretation comprehensible.

Students' exploration behaviour

In order to gather knowledge about the effects of a creature's genes on its characteristics, students have to explore it by actively manipulating the genes (see Figure 10.1). These manipulations, however, are most informative if students switch one gene to "on" and all other genes to "off". Only then can they unambiguously interpret changes in the creature's characteristics as effects of the gene that is switched "on" (Vollmeyer, Burns and Holyoak, 1996). Moreover, in order to detect dynamics within a creature (i.e. characteristics which change without being affected by genes), all genes have to be switched "off". Thus, when looking at the students' traces, such informative steps can be distinguished from non-informative ones. This behavioural information can then be used to relate the number of informative steps to the total number of steps taken in the exploration phase across all creatures, resulting in a behaviour-based systematic exploration (SE) score (Kröner et al., 2005). A high proportion of informative steps indicates a very efficient exploration strategy – the student in question mostly applies informative steps.

As can be seen in Table 10.2, internal consistency on this performance score is very high, ranging from .88 to .94. The score's validity is further supported by its substantial correlation to the students' gathered system knowledge (SK) score. The more efficiently students explore creatures, the greater their knowledge about them. Interpretation of the score is rather easy, with a theoretical range of 0 to 100, with 100 indicating that the student has only applied informative steps.

To ease and support its use in educational practice, the GL package (available under an open-source licence at www.assessment.lu/GeneticsLab) provides additional information about the student's exploration behaviour. Besides the efficiency of the exploration strategy applied, educators can determine whether the effects of all genes have been investigated, that is, whether the student has realised all possible informative steps. This may be valuable information for interventions with students who explore highly efficiently but not conscientiously.

Students' gathered knowledge

The knowledge students gathered about a creature is scored on basis of the causal diagrams which were created by the students in the GL database (see Figure 10.2). These causal models can be interpreted as the theoretical model students have developed about the creatures and are thus valid indicators of their mental problem representations (Funke, 1992). Although causal diagrams have often successfully been used for knowledge assessment among samples of university students (Blech and Funke, 2005; Funke, 2001), one critique raised was, that due to its high cognitive demand, this approach might be problematic for assessing students with lower cognitive capacity. This was in fact supported by the results of our usability studies, showing that many students in our target population reported problems using causal diagrams for knowledge representation. Although we tackled this problem by a substantial modification of the related instructions it still had to be confirmed that the analysis of causal diagrams would yield reliable and valid scores of students' problem representation in our target sample. To score the resulting causal diagrams, we applied a well-established algorithm that differentiates between relational knowledge (i.e. whether a relation between a gene and a characteristic exists or not) and knowledge about the strength of these relations (Funke, 1992; Müller, 1993). The student's model is compared to the true underlying relationships and the more similar they are, the higher the student's knowledge score. Both kinds of knowledge are scored separately and then weighted in order to compose a total system knowledge score. In line with previous studies, relational knowledge was emphasised by multiplying it with a weight of .75, compared to a weight of .25 for knowledge about the strength of an effect (Funke, 1992).

Table 10.2 shows that the resulting score about the students' gathered knowledge is highly reliable, with a Cronbach's alpha ranging from .77 to .90. Descriptives of this score are given as the percentage achieved of a maximum score (POMP, see above). Moreover, the pattern of intercorrelations between system knowledge and systematic exploration as well as control performance supports the internal validity of the score: a more efficient exploration strategy leads to a higher system knowledge score and the higher the gathered knowledge, the better the ability to achieve the target values. Thus, system knowledge can be seen as a reliable and valid measure of students' mental problem representations. In addition to the total system knowledge score, the GL package also provides both specific knowledge scores: relational knowledge and knowledge about the strengths of effects. To ease interpretation of the score for educators, the GL's manual contains theoretical minima as well as maxima for each score.

Students' control performance

To score students' ability to apply their gathered knowledge and achieve certain target values for the creatures' characteristics (see Fig. 10.3), we again drew on behavioural data to compute a process-oriented control performance (CP) score. In order to achieve the target values within three steps, students have to 1) rely on their knowledge to plan their actions and to forecast possible consequences; and 2) react to unexpected consequences and try to correct them. Both skills are key characteristics of CPS (Funke, 2003). Most previous attempts to score control performance emphasised the (aggregated) deviation between the achieved values and the target values (Blech and Funke, 2005). This approach, however, has been criticised for making the scoring of a step dependent on the previous one, if the scenario does not allow students to reach the target value in a single step. A suboptimal step would automatically lead to a deviation from the target value that could not be compensated by the following step. To put it differently, high levels of skill in correcting problem states could not compensate for poor planning behaviour. Consequently, we developed a scoring algorithm that is exclusively based on the students' inputs and that scores every step independently. Only if a step is optimal, in the sense that it maximally decreases the distance to the target values, is it seen as indicating good control performance. Thus, for each creature a maximum score of three is possible.

Internal consistency of the resulting control performance score was generally acceptable (see Table 10.2). However, the value for Cronbach's alpha was rather low for Study 2, indicating that the combination of both interacting of both interacting with the creature, reacting, and correcting current states may complicate the scoring of students' control performance. Nevertheless, the results of Study 3, the most representative study, together with the meaningful pattern of intercorrelations throughout all studies – high system knowledge leads to better control performance – suggest the score's validity. In addition to the number of optimal steps taken by students, educators can also find the concrete sequence of steps they took. Interventions could therefore either target students who lack planning skills given a suboptimal first step, or students who show poor control behaviour.

Table 10.2. **Means, standard deviations, reliability and intercorrelations of the Genetics Lab's performance scores**

	Number of items	Cronbach's alpha	Mean	Standard deviation	Min	Max	25th percentile	Median	75th percentile	Correlation with		
										SE	SK	CP
Study 1 (n = 43) **Complex problem solving**												
Systematic exploration	16	0.94	21	12	1	61	13	21	27	1		
System knowledge	16	0.89	54	12	38	96	46	51	57	0.54	1	
Control performance	16	0.79	32	7	16	47	26	31	35	0.27	0.43	1
Acceptance and usability												
Perceived ease of use	4	0.71	54	23	0	100	44	56	69	0.31	0.44	0.39
Attractivity	9	0.91	64	22	0	100	56	67	78	0.22	0.34	0.54
Comprehensibility	10	0.81	61	17	20	100	50	63	73	0.28	0.49	0.32
Functionality	7	0.82	60	22	0	100	46	64	71	0.17	0.35	0.5
Study 2 (n = 61) **Complex problem solving**												
Systematic exploration	12	0.88	26	11	7	66	19	25	32	1		
System knowledge	12	0.77	53	12	35	100	45	51	59	0.35	1	
Control performance	12	0.54	21	4	11	31	18	21	24	0.32	0.47	1
Study 3 (n = 563) **Complex problem solving**												
Systematic exploration	12	0.91	28	15	1	71	17	26	39	1		
System knowledge	12	0.9	69	17	37	100	55	67	81	0.55	1	
Control performance	12	0.79	20	6.8	6	36	14	18	24	0.51	0.77	1

Note: SE: Systematic exploration; SK: System knowledge; CP: Control performance.

Summary and outlook

The classroom assessment of complex problem solving places special demands on the assessment instruments used for this purpose. The development of the Genetics Lab has successfully responded to most of these challenges by drawing on game-like characteristics, a highly usable user interface and psychometrically sound, behaviour-based scores that are also comprehensive for educators. However, several questions remain to be answered. First, although the vast majority of students in our studies showed characteristics of being "digital native" and thus were likely to be highly competent in using computers and digital media, a minority of students continue to report low ICT competency and only occasional use of modern media. To what extent these students are disadvantaged by the computer-based assessment of their problem-solving skills has to be further investigated. Still, given that most future problems of high complexity will be solved in a digital environment this may not be a shortcoming of the assessment instrument but instead contribute to its external validity.

Second, although the implementation of game-like characteristics in the GL leads to high acceptance among today's students, these features could interfere with the construct being measured by making the problems it presents especially interesting and attractive for some, but not for others. Studies investigating the GL's concurrent validity with other measures of problem solving are therefore needed. Finally, although the scores provided by the GL proved to be internally valid and reliable, their usefulness has yet to be demonstrated in studies that use them for evaluations or interventions. The use of behavioural data is still in its infancy and could substantially benefit from such experiences. In developing the Genetics Lab in four different languages and making it freely accessible online, we have taken a first step towards answering these future challenges.

Notes

1 Visit www.assessment.lu/GeneticsLab to download the GL, and for additional information.

References

AERA, APA and NCME (1999), *Standards for Educational and Psychological Testing* American Educational Research Association, American Psychological Association and the National council on Measurement in Education, Washington, DC.

Bennett, R.E., F. Jenkins, H. Persky and A. Weiss (2003), "Assessing complex problem solving performances", *Assessment in Education: Principles, Policy & Practice*, Vol. 10/3, pp. 347-359.

Bennett, S. and K. Maton (2010), "Beyond the 'digital natives' debate: Towards a more nuanced understanding of students' technology experiences", *Journal of Computer Assisted Learning*, Vol. 26/5, pp. 321-331.

Bennett, S., K. Maton and L. Kervin (2008), "The 'digital natives' debate: A critical review of the evidence", *British Journal of Educational Technology*, Vol. 39/5, pp. 775-786.

Blech, C. and J. Funke (2005), *Dynamis Review: An Overview About Applications of the Dynamis Approach in Cognitive Psychology*, Deutsches Institut für Erwachsenenbildung, www.die-bonn.de/esprid/dokumente/doc-2005/blech05_01.pdf.

Cohen, P., J. Cohen, L.S. Aiken and S.G. West (1999), "The problem of units and the circumstance for POMP", *Multivariate Behavioral Research*, Vol. 34/3, pp. 315-346.

Fulcher, G. (2003), "Interface design in computer-based language testing", *Language Testing*, Vol. 20/4, pp. 384-408.

Funke, J. (2003), *Problemlösendes Denken* [Problem-Solving Thinking], Kohlhammer, Stuttgart.

Funke, J. (2001), "Dynamic systems as tools for analysing human judgement", *Thinking and Reasoning*, Vol. 7/1, pp. 69-89, http://dx.doi.org/10.1080/13546780042000046.

Funke, J. (1992), *Wissen über dynamische Systeme: Erwerb, Repräsentation und Anwendung* [Knowledge About Dynamic Systems: Acquisition, Representation and Application], Springer: Heidelberg.

Funke, J. and P.A. Frensch (2007), "Complex problem solving: The European perspective – 10 years after", in D.H. Jonassen (ed.), *Learning to Solve Complex Scientific Problems*, Erlbaum, New York, pp. 25-47.

Greiff, S., S. Wüstenberg, S and J. Funke (2012), "Dynamic problem solving: A new assessment perspective", *Applied Psychological Measurement*, Vol. 36/3, pp. 189-213, http://dx.doi.org/10.1177/0146621612439620.

Hadwin, A. F., J.C. Nesbit, D. Jamieson-Noel, J. Code and P.H. Winne (2007), "Examining trace data to explore self-regulated learning", *Metacognition and Learning*, Vol. 2/2, pp. 107-124.

Jones, C., R. Ramanau, S. Cross and G. Healing (2010), "Net generation or Digital Natives: Is there a distinct new generation entering university?", *Computers & Education*, Vol. 54/3, pp. 722-732.

Kröner, S., J.L. Plass and D. Leutner (2005), "Intelligence assessment with computer simulations", *Intelligence*, Vol. 33/4, pp. 347-368.

Leutner, D., J. Fleischer, J. Wirth, S. Greiff and J. Funke (2012), "Analytische und dynamische Problemlösekompetenz im Lichte internationaler Schulleistungsvergleichsstudien" [Analytic and dynamic problem solving competence in the light of international comparative student assessment studies], *Psychologische Rundschau*, Vol. 63/1, pp. 34-42.

Li, Y. and M. Ranieri (2010), "Are 'digital natives' really digitally competent? A study on Chinese teenagers", *British Journal of Educational Technology*, Vol. 41/6, pp. 1029-1042.

Margaryan, A., A. Littlejohn and G. Vojt (2011), "Are digital natives a myth or reality? University students' use of digital technologies", *Computers & Education*, Vol. 56/2, pp. 429-440.

Mayer, R.E. (2003), "The promise of multimedia learning: Using the same instructional design methods across different media", *Learning and instruction*, Vol. 13/2, pp. 125-139.

McPherson, J. and N.R. Burns (2007), "Gs Invaders: Assessing a computer game-like test of processing speed", *Behavior Research Methods*, Vol. 39/4, pp. 876-883.

Müller, H. (1993), *Komplexes Problemlösen: Reliabilität und Wissen* [Complex Problem Solving: Reliability and Knowledge], Holos, Bonn, Germany.

OECD (2013), *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*, OECD Publishing, Paris.

OECD (2004), *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*, OECD Publishing, Paris.

Prensky, M. (2001a), "Digital natives, digital immigrants part 1", *On the Horizon*, Vol. 9/5, pp. 1-6.

Prensky, M. (2001b), "Digital natives, digital immigrants part 2: Do they really think differently?", *On the Horizon*, Vol. 9/6, pp. 1-6.

Ridgway, J. and S. McCusker (2003), "Using computers to assess new educational goals", *Assessment in Education: Principles, Policy & Practice*, Vol. 10/3, pp. 309-328.

Rigas, G., E. Carling and B. Brehmer (2002), "Reliability and validity of performance measures in microworlds", *Intelligence*, Vol. 30/5, pp. 463-480.

Rollett, W. (2008), *Strategieeinsatz, erzeugte Information und Informationsnutzung bei der Exploration und Steuerung komplexer dynamischer Systeme* [Strategy Use, Generated Information and Use of Information in Exploring and Controlling Complex, Dynamic Systems], Lit Verlag, Berlin.

Shum, D., J. O'Gorman and B. Myors (2006), *Psychological Testing and Assessment,* Oxford University Press, Melbourne.

Sonnleitner, P. et al. (2012a), "The Genetics Lab: Acceptance and psychometric characteristics of a computer-based microworld assessing complex problem solving", *Psychological Test and Assessment Modeling*, Vol. 54/1, pp. 54-72.

Sonnleitner, P. et al. (2012b), *Genetics Lab: A Computer-Based Microworld to Assess Complex Problem Solving: Theoretical Background & Psychometric Evaluation*, Research Report, University of Luxembourg.

Sonnleitner, et al. (2013). Students' complex problem-solving abilities: Their structure and relations to reasoning ability and educational success. *Intelligence*, 41, 289–305.

Süß, H.M. (1996), *Intelligenz, Wissen und Problemlösen: kognitive voraussetzungen für erfolgreiches Handeln bei computersimulierten Problemen* [Intelligence, Knowledge and Problem Solving: Cognitive Prerequisites for Success in Problem Solving with Computer-Simulated Problems], Hogrefe, Göttingen.

Tapscott, D. (1998), *Growing up Digital: The Rise of the Net Generation*, McGraw-Hill, New York.

Terzis, V. and A. Economides (2011), "The acceptance and use of computer based assessment", *Computers & Education*, Vol. 56/4, pp. 1032-1044.

Veen, W. and B. Vrakking (2006), *Homo Zappiens: Growing Up in a Digital Age*, Network Continuum Education, London.

Vollmeyer, R., B.D. Burns and K.J. Holyoak (1996), "The impact of goal specifity on strategy use and the acquisition of problem structure", *Cognitive Science*, Vol. 20/1, pp. 75-100.

Washburn, D.A. (2003), "The games psychologists play (and the data they provide)", *Behavior Research Methods*, Vol. 35/2, pp. 185-193.

Weiler, A. (2005), "Information-seeking behavior in Generation Y students: Motivation, critical thinking, and learning theory", *The Journal of Academic Librarianship*, Vol. 31/1, pp. 46-53.

Winne, P.H. (2010), "Improving measurements of self-regulated learning", *Educational Psychologist*, Vol. 45/4, pp. 267-276.

Wirth, J. and E. Klieme (2003), "Computer-based assessment of problem solving competence", *Assessment in Education: Principles, Policy & Practice*, Vol. 10/3, pp. 329-345.

Wood, R.T., M.D. Griffiths, D. Chappell and M.N. Davies (2004), "The structural characteristics of video games: A psycho-structural analysis", *CyberPsychology & Behavior*, Vol. 7/1, pp. 1-10.

PART IV

New indicators

Chapter 11

Log-file data as indicators for problem-solving processes

By

Nathan Zoanetti

Victorian Curriculum and Assessment Authority, Melbourne, Australia.

and Patrick Griffin

Assessment Research Centre, University of Melbourne, Australia.

This chapter describes a theoretical rationale and an empirical methodology for analysing log-file data recorded from students' interactions with computer-based problem-solving assessment tasks. These analyses can play an important role in the development and refinement of rules for automatically scoring complicated sequences of process data which describe procedural aspects of problem solving. It presents examples of log files and analyses from an interactive problem-solving task. In complex domains where a variety of skills and dispositions may influence performance, interpreting log-file data to infer how students solve problems could provide valid information to help meet their instructional needs.

This research was supported by an Australian Research Council Linkage Grant (LP0561852) in partnership with North Shore Development College Pty Ltd. The views expressed by N. Zoanetti in this chapter are those of the author and do not necessarily reflect the views of the Victorian Curriculum and Assessment Authority.

Introduction

The assessment of problem-solving skills using interactive, computer-based tasks is gaining momentum (Richardson et al., 2002; Bennett et al., 2003; Wirth and Klieme, 2003; Zoanetti, 2010; Greiff, Wüstenberg and Funke, 2012; Griffin et al., 2013). One of the primary challenges is making sense of the large amounts of data elicited and captured by such assessments. These data typically describe student-task interactions and, while they are readily available, their interpretation and psychometric application is no simple matter (Williamson, Bejar and Mislevy, 2006). This research area requires further development to better support the measurement and teaching of problem-solving skills. The chapter presents a theoretical rationale and an empirical methodology for the analysis of log-file data in problem-solving assessments.

What are log files?

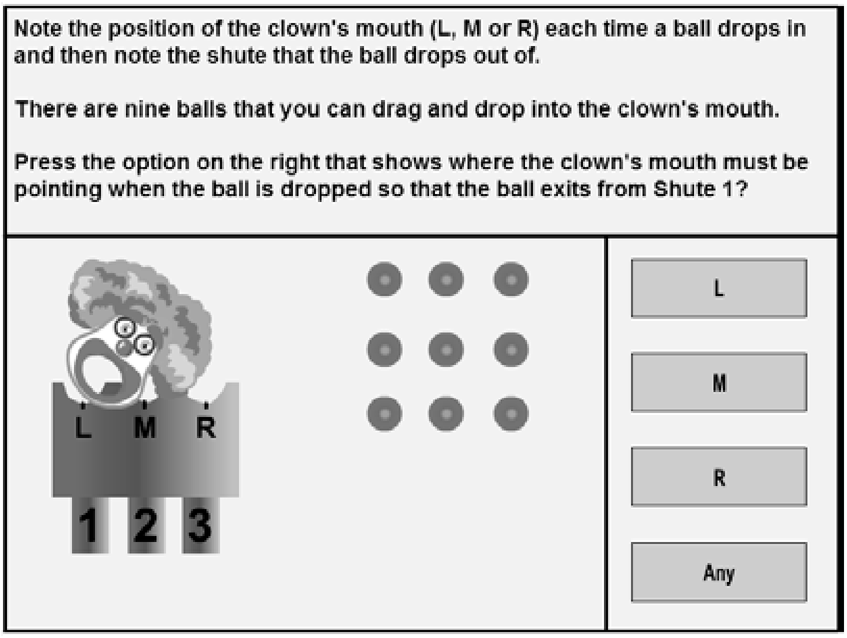
In the context of computer-based assessment, student-task interactions are easily recorded to produce what is often called a log file. These records typically describe distinct keystrokes and mouse events (typing, clicking, dragging, dropping, cursor movement, etc.). Each discrete action would usually be recorded with a corresponding time stamp, or time of occurrence. These time-stamped interaction data have been referred to variously as "log-file data" (Arroyo and Woolf, 2005), "discrete action protocols" (Fu, 2001), "click-stream data" (Chung et al., 2002) and "process data" (Zoanetti, 2010). Other variants no doubt exist.

Log files can easily be generated and stored as plain text files, often in the form of delimited strings of text. Alternatively, more formal database architectures can be used as in the project described by Care and Griffin (Chapter 14; see also Chapter 12 for log-file analyses). Figure 11.2 presents an example of a log file, in the form of delimited strings in a text file. To aid interpretation, the task from which these data were recorded is shown in Figure 11.1.

The main features of the record in Figure 11.2 are delimited by the colon character. The first value of "99999999" represents the student identifier (de-identified here); the second value of "20071025045942" is a date and time stamp for when the record was logged; the third value is the item identifier "0101" prefixed with "it_"; the fourth value is the final result or product submitted by the student (which in this case corresponds to the chosen option "R" in Figure 11.1), prefixed with "res_"; the fifth value is the total response time of "67.706" seconds for the task, prefixed with "rt_"; and the final value is the sequence of process data recorded by the task interface as the student worked through the problem, prefixed with "sq_". This sequence records discrete mouse cursor events and corresponding data along with their time of occurrence in seconds since the task was rendered on screen.

For example, the first element corresponds to the student commencing dragging the "11" ball after 23 seconds, which is the top left ball of the nine balls seen in Figure 11.1. The second element in the sequence, which occurred after 26 seconds, corresponds to the student dropping the "11" ball. Following this element is a secondary element which specifies the x- and y-coordinates of the location where the ball was dropped. In this case the x-coordinate was 115.55 and the y-coordinate was 111.35. Other mouse events such as rolling of the cursor over the options buttons (e.g., "roL", "roM" and "roR") can also be found within the process data sequence. As is evident in Figure 11.2, even for a relatively simple task, the recorded actions can quickly culminate in a large amount of raw data describing student problem-solving behaviour. Importantly, patterns of behaviour can be linked to theories of cognition and developing competence (Pelligrino, Chudowsky and Glaser, 2001; Williamson, Bejar and Mislevy, 2006). Some relevant problem-solving theories are discussed in the following section.

Figure 11.1. **Interactive Laughing Clowns task (the clown's mouth is in constant motion)**

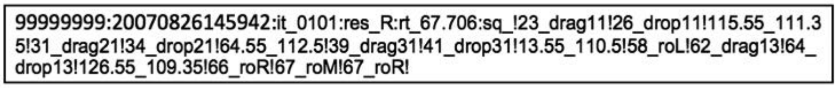


Note the position of the clown's mouth (L, M or R) each time a ball drops in and then note the shute that the ball drops out of.

There are nine balls that you can drag and drop into the clown's mouth.

Press the option on the right that shows where the clown's mouth must be pointing when the ball is dropped so that the ball exits from Shute 1?

Figure 11.2. **Sample log-file record from a text file recorded by the interactive Laughing Clowns task**



```
99999999:20070826145942:it_0101:res_R:rt_67.706:sq_!23_drag11!26_drop11!115.55_111.3
5!31_drag21!34_drop21!64.55_112.5!39_drag31!41_drop31!13.55_110.5!58_roL!62_drag13!64_
drop13!126.55_109.35!66_roR!67_roM!67_roR!
```

A theoretical rationale for analysing log files in problem-solving assessments

Problem solving as sequential steps

Problem solving is commonly defined as the application of cognitive processes to reach a goal where there is no obvious means for doing so (Mayer, 1982). A problem is considered to have an initial state, one or more goal states, and some number of intermediate states (Newell and Simon, 1972). It follows that problem solving can be conceptualised as a sequential process where the problem solver must understand the problem, devise a plan, carry out the plan, and monitor their progress in relation to the goal (Garofalo and Lester, 1985; Polya, 1945, 1957). This definition of problem solving as sequential phases has implications for the kinds of assessment inferences one might like to make. For instance, it could be of value to determine how students approach the early stages of solving a problem: do they rush to interact with the problem? Do they spend time reading and planning first? Do they tend to act on a problem only when they understand it? Such questions refer to performance in the initial phases of problem solving, and make no reference to whether the student eventually solves the problem. It is also possible to move the target inferences to the intermediate stages of the problem (how systematically does the student work through problems?) or to the final stages of the problem (does the student tend to check their solution?). In other words, it seems reasonable that inferences might be made about distinct procedural aspects of problem solving. If a student's strengths, weaknesses and specific strategies can be described at different stages of problem solving, then a rich and relatively fine-grained diagnosis of student problem-solving proficiency could be made available to inform targeted instructional interventions. With these potentially useful inferences in mind, the challenge for assessment designers is to determine the kinds of evidence and corresponding tasks required to reveal them.

Interactive problem solving and hypothetico-deductive reasoning

Many real-life problems cannot be solved without experimental interaction (Frensch and Funke, 1995; Leutner and Wirth, 2005; Wirth and Klieme, 2003). This arguably grants interactive problem solving considerable educational importance. Naturally, when an important ability is identified, there will be a desire to model, measure and infer how to teach the competencies associated with that ability. The need for interactive engagement with real-life problems suggests that assessing performance on interactive tasks would be useful if the information gained from doing so could be used to improve the associated problem-solving skills. The following sections examine the extent to which data from interactive tasks can be related to theories of development and expertise in problem solving.

It is also possible to conceptualise problem-solving proficiency on interactive tasks in accordance with a developmental framework much like the one described by Callingham and Griffin (2000). They helped to point to a way of developing ordered categories of response quality for various investigative mathematical problem-solving tasks. In the first and lowest category (beyond random guessing), students were shown to rely on identifying isolated elements of information, which become patterns of observations. At the next level of problem analysis, students formed and used rules, either to regulate the task or to monitor and reflect via hypothesis testing. At a more sophisticated level, students used rules to complete steps of the problem. For the most difficult subtasks or aspects of tasks, more able students demonstrated an ability to generalise to a range of situations by setting and testing hypotheses using a "what if…?" approach. Callingham and Griffin (2001) thus proposed a progression of pattern, rule, generalisation and hypothesis.

For a developmental perspective on problem solving, we need clear definitions of what students are expected to learn, and a theoretical framework of how that learning is expected to unfold as the student progresses through the instructional material. In particular, there is a need to describe what behaviour might be observed in students who possess different amounts of the relevant problem-solving construct. This has important implications for assessment design and scoring because it highlights the kinds of situations needed to gather evidence of differences between students, and also the value of certain observable behaviour as indicators of development.

The developmental framework described in the previous paragraphs is partly based on a progression of general skills in inductive and deductive reasoning. While inductive reasoning focuses on establishing a possible explanation to test in the first place, deductive reasoning involves testing whether the explanation is valid or not. This combination is known as hypothetico-deductive reasoning and is often taken to be the defining characteristic of scientific practice. An important step in this method is the generation of hypotheses based on observations and "speculating" or posing questions. The next step is to make observations and gather data that allow those hypotheses to be tested. The deductive method attempts to "deduce" facts by eliminating all possible outcomes that do not fit the facts. Interactive problem-solving domains typically require that students reach a conclusion on the nature of hypotheses to be tested and the manner in which these hypotheses will be tested. Log-file data naturally embed the breadth and depth of interaction information required for monitoring the kinds of tests students carry out, including their number, sequencing and timing.

An inductive approach adopts the view that if certain data or observations continually attain a consistent result, then that result must be true. It was this kind of thinking that led early scientists to believe in spontaneous generation. If a result does not fail or until someone challenges the generalisation, the observation is regarded as "true". This suggests that the students need encouragement to explore whether there might be an alternative explanation, or whether the generalisation stands. By examining the problem space carefully and insightfully, by being allowed to be puzzled by phenomena that they think they understand, by being allowed to "unknow" what they know, it is likely that students will take a first step in their knowledge development towards finding and asking questions that can lead to a theory. This framework reveals the nature of reasoning underlying certain aspects of interactive problem solving. It reveals the nature of trial-and-error sequences, and monitoring and reflection, that students might undertake to move from random trial and error through to higher level hypothesis testing and reflection. These substantively different developmental phases can be used as a reference point from which to develop rules for scoring rich sequences of log-file data, where each score might unambiguously indicate a distinct level of development.

Using theories of expertise to determine the indicators for expert problem solving

Whether or not a problem is successfully solved can be informative about the proficiencies of a problem solver. However, where multiple strategies might lead to a solution, or where certain strategies might be applied with greater or less efficacy, it may be beneficial to probe the specific processes undertaken by the problem solver in more detail. Inferences relating to the choice of strategy and the quality of how it is applied then dictate the types of evidence required from task-design efforts. A set of relevant indicators can be specified based on theories of expertise in problem solving.

Differentiating between expert and novice performances can reveal the key indicators of cognition which characterise higher levels of problem-solving performance in a given domain. As Glaser stated, "…studies of expertise have investigated the nature of the knowledge and cognitive processes that underlie developing competence in various domains of learning" (1991:17). A prior knowledge of these indicators is crucial for assessment design, since these indicators suggest the range of observations that assessment tasks should be designed to elicit to make it possible to differentiate between the performance levels of problem solvers (Mislevy, 2008; Pelligrino, Chudowsky and Glaser, 2001). A number of well-documented theories of problem-solving expertise provide inventories of the types of observations which might help discriminate between expert and novice problem solvers. As an example, Glaser (1991) stated that while experts tend to solve problems faster than novices, they often spend disproportionately longer decoding problems before acting on them. The explanation is that experts place greater emphasis on planning and analysis in the initial stages of problem solving, and are less likely to be distracted by, and interact with, superfluous problem features (Chi, Feltovich and Glaser, 1981; Sternberg and Ben-Zeev, 2001). Empirical support for these theories has been gathered by Larkin and colleagues (Larkin et al., 1980) who found that experts would devote a disproportionate amount of time reflecting on the nature of physics problems prior to engaging in solution steps. A similar finding is reported by Paek (2002) in the context of mathematics problems, where students who successfully solved problems were found on average to spend 35% more time on the first step of the problem. This brief discussion of expert-novice differences already highlights several performance features and corresponding data which might be informative, including the time students spend working on tasks, the time students spend before they interact with tasks, the quality of their initial interactions with tasks, the total number of interactions with tasks and whether or not they successfully complete tasks. These data tend to be readily available in log files which include a time stamp for each student-task interaction.

While there are several general theories of expert-novice differences in problem solving (Glaser, 1991; Mayer, 1982; Newell and Simon, 1972), more specialised conceptualisations have recently emerged in technology-rich, interactive problem-solving domains. Two examples include the work of Stevens and colleagues (Stevens and Thadani, 2006; Thadani, Stevens and Tao, 2009) on interactive chemistry problems and the work of Wirth and colleagues on finite state automata tasks (Leutner and Wirth, 2005; Wirth and Klieme, 2003). The primary indicators used to differentiate between problem solvers in these complex and interactive task domains were notions of efficiency and automaticity. Efficiency can be inferred from indicators such as the number of actions taken to solve a problem and the tendency of a problem solver to avoid errors and repetition (Newell and Simon, 1972). Automaticity can be related to the amount of time between actions and the total response time (Gitomer and Rock, 1993; Mislevy, 1993). In summary, problem solvers' overall expertise can be inferred from conclusions about their goal attainment, efficiency and automaticity (Lohman and Ippel, 1993:56; Quellmalz and Pellegrino, 2009). These concepts again provide the basis for proposing the structure and type of data that should be collected and evaluated by assessment systems designed to measure proficiency in interactive problem solving.

Based on the discussion above, it is possible to specify an inventory of performance features which would be expected to differentiate between relatively novice and expert solvers in the interactive problem-solving domain. The programme of problem-solving research which led to the construction of the task in Figure 11.1 took into consideration the following observable features of performance data which could be logged directly, or extracted from log files with varying degrees of objectivity: action count, error count, invalid action count, repeated action count, total response time, latencies between key events, time to first action, search scope or count of distinct problem states reached, correctness of outcome, count of departures from the goal state, count of visits to the goal state, action rapidity, task-specific subgoals attained, task-specific action sequences, quality of initial actions and validity of initial actions.

Rules for extracting values for these observables from log-file data can be developed, refined and validated through empirical methodologies such as cognitive task analyses (Mislevy et al., 1999). These rules could be relatively straightforward. For example in Figure 11.2, for the "correctness of outcome" indicator, a final value of "R" would be assigned a score of one and all other values would score zero. More complicated but still reasonably objective rules for the task in Figure 11.1 could include the scoring of trial-and-error sequences such as whether the scope of a student's exploration of the problem space was adequate. If we envisage a partial credit scoring scenario, students might receive zero marks for not exploring the system, partial marks for incomplete exploration and full marks for complete exploration of the system. A category could even be included to capture whether students "over-explored" the system, exhibiting a great deal of redundancy. Developing a corresponding automated scoring algorithm for the task in Figure 11.1 and the log-file data in Figure 11.2 might first involve processing the log-file data to extract the sequence of positions (which we can refer to as the "drop sequence") where the student dropped the ball. In the case of Figure 11.2, this would produce the string "RMLR" since the student in fact dropped the balls into the clown mouth positions sequentially from right to middle to left, and then, for reasons elaborated in the discussion of Figure 11.3, conducted one further test of position "R". A scoring rule can be qualitatively described and specified for automation as in Table 11.1.

The student whose log file is shown in Figure 11.2, would score 3 for this indicator since their trial-and-error sequence started with "RML" and was followed by one further trial of "R". This rule contains certain embedded assumptions, and might undergo refinement as exceptions are observed during cognitive task analysis and wider task piloting. For instance, it might be that the score of 2 poorly discriminates between students at different levels of the construct. If that were the case, the rule might be refined to omit this category and relax the constraint in the highest category so that it is awarded irrespective of the number of balls used as long as at least one trial of each of "L", "M" and "R" was conducted. These sorts of decisions are best left as empirical questions, with processes in place to ensure that student data can be used to inform refinements to optimise the reliability and validity of the assessment by minimising sources of construct-irrelevant variance and resulting measurement error.

For some indicators, the provisional or theorised rules may lead to much less objective scoring and would need continuous tuning, such as when the timing of specific sequences of actions might be evaluated to determine if a specific strategy has been applied. Subsequent sections of this chapter discuss a rule-refinement and validation methodology from a study of 8-12 year-old Australian school students.

Table 11.1. **A possible partial credit framework for scoring an indicator describing the quality of exploration**

Score	Preliminary scoring rules	Dominant theoretical explanation	Possible competing explanations
0	The drop sequence is null, i.e. the student did not conduct any trials.	Inadequate problem representation.	Insufficient motivation.
1	The drop sequence doesn't contain each of "L", "M" and "R" at least once, i.e. at least one position was not explored/eliminated.	Inadequate problem representation; inadequate monitoring.	Student satisfied with result having identified both the "correct" pipe position and one "incorrect" position.
2	The drop sequence contains three of each "L", "M" and "R", i.e. the student used all of the available balls.	Novice problem solvers exhibit greater redundancy in their exploration of the problem space.	Student completed the remaining trials for reasons unrelated to the construct, e.g. for fun.
3	The drop sequence contains at least one each of "L", "M" and "R" but some balls remained unused.	Experts tend to have deep representations of problems and are efficient in exploring the problem space.	

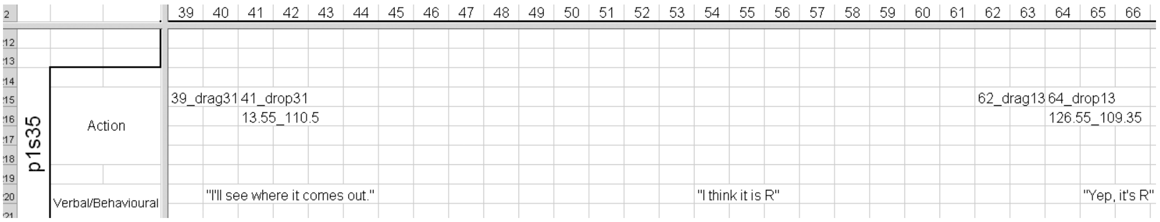
Analysis of log files

It is possible to analyse log files in several ways, for a variety of purposes. For interactive tasks and their corresponding log-file data to be part of an efficient assessment format, "…it will be necessary for the computer to be responsible not only for item administration but also for item scoring" (Margolis and Clauser, 2006:164). It is important to establish reliable scoring rules which can be programmed and applied to log-file data. Williamson, Bejar and Mislevy explain that a "…key to leveraging the expanded capability to collect and record data from complex assessment tasks is implementing automated scoring algorithms to interpret data of the quantity and complexity that can now be collected" (2006: 2). Some methods call on machine learning algorithms (Chapter 12). Other methods rely on identifying specified performance indicators in advance on a theoretical and perhaps empirical basis. In this latter approach, specific features in the process data are evaluated or scored using rule-based algorithms (Braun, Bejar and Williamson, 2006). These have been referred to as "evidence rules" by Mislevy and colleagues (2006). Cognitive task analysis methodologies can be informative for ensuring that tasks elicit the anticipated range of behaviour and corresponding data (Williamson et al., 2004), and for establishing and refining evidence rules, as discussed in the following section.

Combining evidence to understand log-file data

Verbal data can be captured by asking students to "think aloud" as they work through problems, and recording their utterances as they go (Ericsson and Simon, 1993). As discussed above, time-stamped computer-captured data are readily available in typical log files. Several assessment research programmes have combined multiple sources of evidence, including verbal and computer-captured data, for the purpose of validating cognitive processes (Chung et al., 2002; Goldman et al., 1999; Van Gog et al., 2005). These different forms of evidence would not necessarily share the same sources of error and incompleteness, and can thereby expand the available data about actual problem-solving strategies (Garner, 1988:70). Goldman and colleagues (1999) were able to show that expert raters using both sources of evidence were able to interpret student solution paths whereas raters relying on click-stream data alone were not. This is an important result which indicates that the reasoning behind complicated log-file sequences (such as systematic trial-and-error sequences) might best be discerned or confirmed by matching the time of their occurrence with corresponding verbal reports. Verbal protocols can therefore be used to “calibrate” interpretations drawn from the log-file data and there is also thought to be some benefit in calibrating in the other direction. It follows that this information could be used to develop and refine rules and algorithms for automatic scoring. Similar concurrent evidence methodologies have proven fruitful before (Garner et al., 1984; Phifer and Glover, 1982). All of these researchers tabulated the two sources of data in the following way: they recorded each computer-captured datum with a corresponding latency (time of occurrence) and, where available, a summary of the corresponding component of the verbal report. To reproduce more authentically the temporal characteristics of students’ problem-solving solution paths, Zoanetti (2010) expanded upon the usual tabulation method by mapping the sources of convergent data onto a horizontal time axis with gradations of one second. Sources of different but concurrent evidence were aligned vertically. Sequences of data from different problem solvers were displayed one on top of another. The resulting maps provided a comprehensive visual depiction of each problem-solving solution path by displaying audio, observational and computer-captured data recordings as they occurred. This analytical tool was subsequently named a temporal evidence map. Examples of temporal evidence map segments, and possible interpretations, are presented in the sections that follow.

Figure 11.3. **Temporal evidence map segment illustrating hypothetico-deductive reasoning from the single-player Laughing Clowns task**

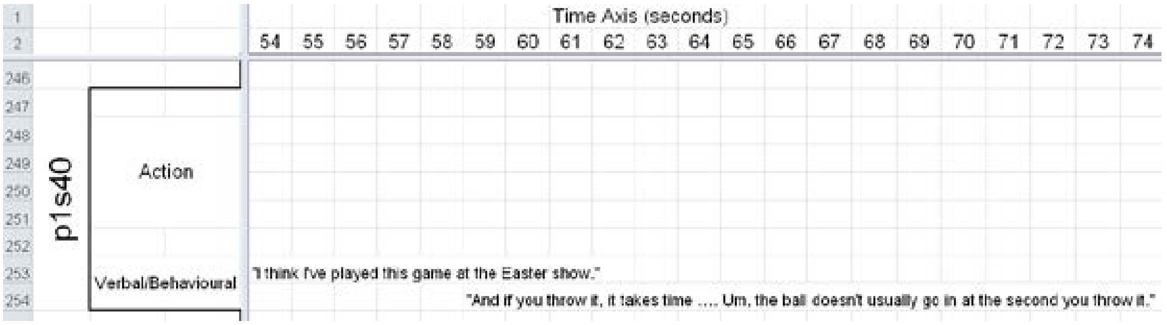


		39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66
p1s35	Action	39_drag31		41_drop31 13.55_110.5																					62_drag13		64_drop13 126.55_109.35		
	Verbal/Behavioural		"I'll see where it comes out."														"I think it is R"											"Yep, it's R"	

What we see from the student’s utterances in Figure 11.3 is how the segments of process data correspond to the student’s intended strategies and cognitive processing. The student was attempting to determine which clown mouth position to drop the ball into so that the ball exited from pipe 1 (see Figure 11.1). They were testing each of the mouth positions in sequence to identify which one would produce the desired result. In other words, the student carried out systematic exploration of the system. They verbally declared and physically conducted their third sequential drag-and-drop test at 39 through 41 seconds, determining that “R” may indeed be the appropriate position. Their subsequent use of the phrase “I think” perhaps indicates that they were not completely certain. Their conduct of an additional test of the “R” position at 62 through 64 seconds, along with the corresponding confirmatory comment, nicely describes the problem-solving processes undertaken by this student. Whether this description could have been derived on the basis of process data alone is not obvious. The verbal data provide a means for triangulating interpretations of process data against concurrent evidence. It also builds on our understanding of how patterns of process data might correspond to important indicators. For example, in this instance we see that the one final test of the correct option prior to submitting their response could be taken as a confirmatory trial. This information lends itself to the construction of programmable rules for automatically scoring the process data.

In Figure 11.4 we see evidence of a well-documented novice characteristic. The problem solver spends a reasonably lengthy period representing the problem using personal experience rather than exploring the system. The verbal evidence helps to explain one cause of lengthy “time to first action” values. It should be noted that there are other, competing explanations for lengthy “time to first action” values, such as the extended planning time undertaken by experts. It may be the case that these analyses could empirically discover that experts find much shorter periods of time adequate for formulating a deep representation of this relatively simple task. This further illustrates the usefulness of the concurrent evidence approach, since identifying competing explanations for certain observations provides some insight into the sources of measurement error which can detract from the reliability of scoring processes within the assessment system. Table 11.2 presents an example of an initial evidence rule framework for this indicator. What is perhaps less clear in Table 11.2 is whether an ordinal scoring framework is tractable, or whether the rules simply produce nominal categories rather than ordered scores. This could have implications for the kinds of statistical models employed to support measurement in this context.

Figure 11.4. **Temporal evidence map segment illustrating typical exploratory behaviour by novice problem solvers from the single-player Laughing Clowns task**



"I think I've played this game at the Easter show."

"And if you throw it, it takes time …. Um, the ball doesn't usually go in at the second you throw it."

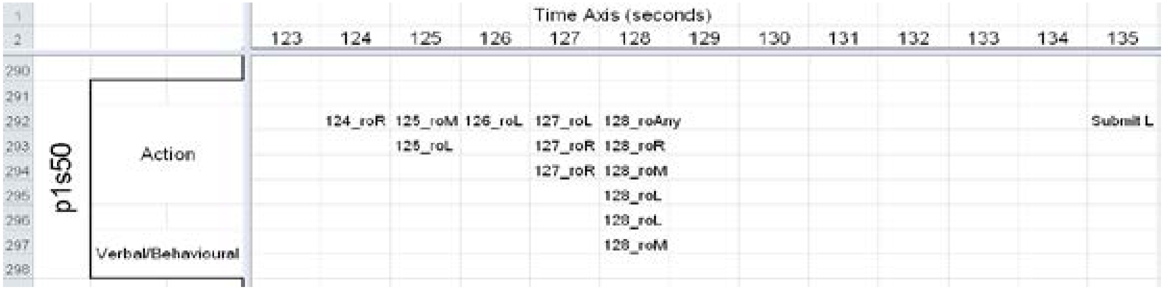
Table 11.2. **Example of a tuneable scoring rule for the "time to first action" indicator**

Preliminary scoring rules	*Dominant theoretical explanation*	*Possible competing explanations*
time to first action <20 seconds: "inadequate"	Exploration prioritised over reading/analysis/planning (novice).	Atypical familiarity with the task (expert, contextual advantage).
time to first action >40 seconds: "excessive"	Excessive delay in or general lack of engagement with task (novice, lack of motivation).	Cautious, extensive planning (expert, cautious disposition).
20 seconds < time to first action <40 seconds: "intermediate"	Reasonable given task reading load and complexity (expert).	

These rules draw on a combination of theory and qualitative data analysis with eventual tuning of the thresholds via more quantitative approaches. While some of these rules were outlined in advance, empirical data were deemed necessary to enhance the level of detail and to maximise the discrimination each indicator contributed towards differentiating between students of different proficiency. Extending the evidence rules in this way is consistent with the work of Bennett and colleagues (2003) on examining problem solving in technology-rich environments, who collected data in an ongoing programme of research to iteratively refine their evaluative scoring rules. This evidence-based rule development becomes crucial for the reliable automation of scoring and provision of feedback.

The competing explanations in the third column of Table 11.2 again deserve comment. If these are not incorporated into the cognitive or analytical models for the assessment, then they could be a source of measurement error and a threat to the reliability of inferences. If it is feasible to disentangle which explanation is at play for a particular student via additional indicators, then this should be pursued and modelled to the extent possible. Where an indicator is likely to be too unreliable, based on the presence of competing explanations identified in theory or through cognitive laboratories, then it should arguably be refined or abandoned in favour of more objective alternatives.

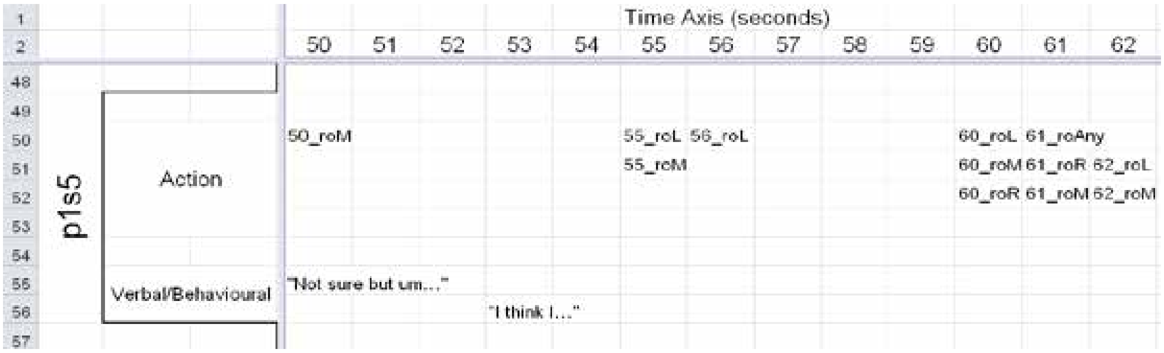
Figure 11.5. **Temporal evidence map segment illustrating guessing behavior from the single-player Laughing Clowns task**



1								Time Axis (seconds)							
2			123	124	125	126	127	128	129	130	131	132	133	134	135
290															
291	p1s50	Action													
292				124_roR	125_roM	126_roL	127_roL	128_roAny							Submit L
293					125_roL		127_roR	128_roR							
294							127_roR	128_roM							
295								128_roL							
296								128_roL							
297		Verbal/Behavioural						128_roM							
298															

Figure 11.5 shows a more subtle indicator of problem-solving behaviour, where the student rapidly rolls the cursor over the various response option buttons. Seven seconds later, they submit an incorrect response, option "L". This example suggests, albeit somewhat subjectively, that rapid rollover sequences could indicate that the problem solver is uncertain about the problem or the solution. This could have implications for scoring the correctness of the outcome, and rules could be evaluated where students with the correct response following a rapid rollover sequence are suspected of guessing the correct answer, or at least not being as sure of the correct answer as students without the rapid rollover sequence. Whether this reasoning could be supported by this arguably weak evidence is perhaps best left as an empirical question. A similar pattern was observed from another student working through the same task, but during the intermediate phases of the solution attempt (see Figure 11.6). Eventually this student attained the correct answer, but their verbal statements and their cursor movements revealed that they were uncertain about the problem at least for the first minute of engagement with it. This in itself may be a useful indicator which lends itself to inferences about the quality of a student's problem representation at various stages during problem solving. These weaker or more subjective indicators might for example have a role in formative assessment contexts, such as prompting clues in intelligent tutoring systems. Machine learning approaches may prove more effective at discerning whether these patterns do constitute useful indicators for discriminating between differentially expert problem solvers.

Figure 11.6. **Temporal evidence map segment illustrating uncertainty about the problem from the single-player Laughing Clowns task**



1									Time Axis (seconds)						
2			50	51	52	53	54	55	56	57	58	59	60	61	62
48															
49	p1s5	Action													
50			50_roM					55_roL	56_roL				60_roL	61_roAny	
51								55_roM					60_roM	61_roR	62_roL
52													60_roR	61_roM	62_roM
53															
54															
55		Verbal/Behavioural	"Not sure but um..."												
56						"I think I..."									
57															

The examples so far reveal examples of behaviour which are from a theoretical standpoint informative for distinguishing between students' competencies, strategies and dispositions. Following scoring of such behaviour via evidence rules, the imperative is to map the patterns of values for an individual student to a generalisable inference about their strengths and weaknesses against prescribed reportable outcome variables. As the discussion section of this chapter will broach, selecting an appropriate statistical model for validly mapping scores to inferences is not necessarily straightforward in the context of complex tasks and correspondingly complex data.

Discussion

Developing interactive tasks and the capacity to automatically score log-file data is a resource-intensive undertaking, even in traditional and well-defined educational domains (Masters, 2010). It is worth questioning whether the educational benefits outweigh the costs, or at least establishing that the educational benefits are clearly apparent to instructors and learners. The key benefits of interpreting and analysing log-file data pertain to the diagnostic richness of the inferences these data can support. A nice way to conceptualise how inferences formulated from process data are likely to boast greater instructional utility than inferences formulated from product data is as follows. Some students may tend to approach problems in an exploratory fashion, learning about the system through a series of rapid, disorganised actions or experiments. Other students may tend to work more methodically and systematically. Yet other students may be unwilling to engage with tasks if they are uncertain about them. A number of general interactive problem-solving typologies have emerged along these lines in studies with interactive problem-solving tasks (Vendlinski and Stevens, 2002; Zoanetti, 2010). These varied strategies may turn out to yield similar numbers of successfully solved problems, suggesting that scoring and scaling the students' answers alone would not be enough to identify and communicate the clear differences between them. Theoretical standpoints might suggest that systematic and efficient routines are superior to rapid and exploratory routines involving more redundancy, but that these latter routines are at least superior to a lack of engagement or experimentation. These differences could have consequences for a learner's ability to successfully engage with complex problems in later educational and vocational settings (Greiff, 2012). An assessment which can identify which "type" of problem solver a student is, in addition to simply assessing whether or not they tend to solve problems, is likely to be more useful for instruction because such an assessment can more explicitly describe learners' strengths and weaknesses as a basis on which to plan intervention.

Large-scale assessment programmes, such as the Programme for International Student Assessment (PISA), customarily apply item response models to report student outcomes on a continuous measurement scale for each educational domain. The extension towards more complex log-file data gathered in less concisely defined educational domains such as interactive problem solving may challenge conventional approaches to scaling educational assessment data (Rupp, 2002). As Bennett et al. (2003) explained, extended performances on complex tasks tend to be multidimensional and elicit data which are locally dependent. Multidimensionality arises when multiple traits of interest influence student performance. Local dependencies can arise when multiple observations are to some extent influenced by being sampled from a common task. Many of the traditional statistical frameworks applied in educational assessment do not adequately handle these complexities. There are exceptions, with item response models being extended to model multiple latent traits (Adams, Wilson and Wang, 1997) and to accommodate local item dependencies (Wainer and Kiely, 1987; Wilson and Adams, 1995). Several researchers have also explored alternative and more flexible approaches to analysing process data in certain problem-solving contexts. For example, artificial neural networks (Vendlinski and Stevens, 2002) were applied to analyse interactive science problems, and Bayesian inference networks have been applied to a range of problem-solving assessments and simulations (Behrens et al., 2004; Bennett et al., 2003; Williamson et al., 2004; Zoanetti, 2010). The axioms of common measurement models, which often do not permit multidimensionality or local item dependencies unless explicitly specified, need to be carefully considered in light of the complexity of the log-file data and the intended granularity of inferences.

Concluding remarks

Log-file data from interactive problem-solving tasks contain patterns of information that are informative about individual students' problem-solving dispositions and abilities. This chapter has detailed a number of documented differences between experts and novices which could be used to compile an inventory of relevant log-file features. Through an analysis of concurrent evidence from the problem-solving process, it should be possible to develop rules to identify and evaluate these salient features. This also serves as an important construct validation phase in the assessment design described in this chapter. Further development and refinement of methodologies for scoring of log-file data to support measurement and instruction in interactive problem-solving domains remains a worthwhile goal in educational assessment research.

References

Adams, R.J., M. Wilson and W.C. Wang (1997), "The multidimensional random coefficients multinomial logit model", *Applied psychological measurement*, Vol.21/1, pp. 1-23.

Arroyo, I. and B.P. Woolf (2005), "Inferring learning and attitudes from a Bayesian network of log file data", in *Proceedings of the 2005 Conference on Artificial Intelligence in Education*, IAED, Amsterdam, pp. 33-40.

Behrens, J. T., et al. (2004), "Introduction to evidence centered design and lessons learned from its application in a global e-learning program", *International Journal of Testing*, Vol. 4/4, pp. 295-301.

Bennett, R.E. et al. (2003), "Assessing complex problem solving performances", *Assessment in Education Principles, Policy and Practice*, Vol. 10/3, pp. 347-359.

Braun, H., I.I. Bejar and D.M. Williamson (2006), "Rule-based methods for automated scoring: Application in a licensing context", in D.M. Williamson, R.J. Mislevy and I.I. Bejar (eds.), *Automated Scoring of Complex Tasks in Computer Based Testing*, Laurence Erlbaum, Mahwah NJ, pp. 83-122.

Callingham, R. and P. Griffin (2001), "Shaping assessment for effective intervention", in *Proceedings of the 18th Biennial Conference of the Australian Association of Mathematics Teachers*, AAMT, Adelaide SA, pp. 271-280.

Callingham, R. and P. Griffin (2000), "Towards a framework for numeracy assessment", in J. Bana and A. Chapman (eds.), *Mathematics Education Beyond 2000 (Proceedings of the 23rd Annual Conference of the Mathematics Education Research Group of Australasia)*, MERGA, Fremantle WA, pp. 134-141.

Chi, M.T.H, P.J. Feltovich and R. Glaser (1981), "Categorization and representation of physics problems by experts and novices", *Cognitive Science,* Vol. 5/2, pp. 121-152.

Chung, G.K.W.K. et al. (2002), "Cognitive process validation of an online problem solving assessment", *Computers in Human Behavior*, Vol. 18/6, pp. 669-684.

Ericsson, K.A. and H.A. Simon (1993), *Protocol Analysis: Verbal Reports as Data*, revised edition, MIT Press, Cambridge, MA.

Frensch, P.A. and J. Funke (1995), "Definitions, traditions, and a general framework for understanding complex problem solving", in P.A. Frensch and J. Funke (eds.), *Complex Problem Solving: The European Perspective*, Erlbaum, Hillsdale, NJ, pp. 3-25.

Fu, W.T. (2001), "ACT-PRO action protocol analyzer: A tool for analyzing discrete action protocols", *Behavior Research Methods, Instruments and Computers*, Vol. 33/2, pp. 149-158, www.ncbi.nlm.nih.gov/pubmed/11447667, accessed 13 December 2012.

Garner, R. (1988), "Verbal report data on cognitive and metacognitive strategies", in C.E. Weinstein, E.T. Goetz and P.A. Alexander (eds.), *Learning and Study Strategies: Issues in Assessment, Instruction and Evaluation*, Academic Press, San Diego CA, pp. 63-76.

Garner, R., V.C. Hare, P. Alexander, J. Haynes and P. Vinograd (1984), "Inducing a look-back strategy among unsuccessful readers", *American Educational Research Journal*, Vol. 21, pp. 789-798.

Garofalo, J. and F. Lester (1985), "Metacognition, cognitive monitoring, and mathematical performance", *Journal for Research in Mathematics Education*, Vol. 16/3, pp. 163-176.

Gitomer, D.H. and D. Rock (1993), "Addressing process variables in test analysis", in N. Frederiksen, R. J. Mislevy and I.I. Bejar (eds.), *Test Theory for a New Generation of Tests*, Laurence Erlbaum, Hillsdale NJ, pp. 243-268.

Glaser, R. (1991), "Expertise and assessment", in M.C. Wittrock and E.L. Baker (eds.), *Testing and Cognition*, Prentice Hall, Englewood Cliffs, NJ, pp. 17-30.

Goldman, S.R., L.K. Zech, G. Biswas, T. Noser and The Cognition and Technology Group at Vanderbilt (1999), "Computer technology and complex problem solving: Issues in the study of complex cognitive activity", *Instructional Sequence*, Vol. 27, pp. 235-268.

Greiff, S. (2012), "Assessment and theory in complex problem solving – A continuing contradiction?", *Australian Journal of Educational and Developmental Psychology*, Vol. 2/1, pp. 49-56.

Greiff, S., S. Wüstenberg and J. Funke (2012), "Dynamic problem solving: A new assessment perspective", *Applied Psychological Measurement*, Vol. 36/3, http://dx.doi.org/10.1177/0146621612439620.

Griffin, P., E. Care, M. Bui and N. Zoanetti (2013), "Development of the assessment design and delivery of collaborative problem solving in the assessment and teaching of 21st century skills project", in E. McKay (ed.), *ePedagogy in Online Learning: New Developments in Web Mediated Human Computer Interaction*, IGI Global, Hershey, PA.

Larkin, J., J. McDermott, D.P. Simon and H.A. Simon (1980), "Expert and novice performance in solving physics problems“, *Science*; Vol. 208/4450, pp. 1335-1342.

Leutner, D. and J. Wirth (2005), "What we have learned from PISA so far: A German educational psychology point of view", *KEDI Journal of Educational Policy*, Vol. 2/2, pp. 39-56.

Lohman, D.F. and M.J. Ippel (1993), "Cognitive diagnosis: From statistically based assessment toward theory-based assessment", in N. Frederiksen, R. J. Mislevy and I. I. Bejar (eds.), *Test Theory for a New Generation of Tests*, Laurence Erlbaum, Hillsdale, NJ, pp. 41-71.

Margolis, M.J. and B.E. Clauser (2006), "A regression-based procedure for automated scoring of a complex medical performance assessment", in I.I. Bejar (ed.), *Automated scoring of complex tasks in computer based testing*, Laurence Erlbaum, Hillsdale, NJ, pp. 123-167.

Masters, J. (2010), "Automated scoring of an interactive geometry item: A proof-of-concept", *Journal of Technology, Learning, and Assessment*, Vol. 8/7, pp. 1-8, http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1626, accessed 14 December 2012.

Mayer, R.E. (1982), "The psychology of mathematical problem solving", in F. Lester and J. Garofalo (eds.), *Mathematical Problem Solving: Issues in Research*, Franklin Institute Press, Philadelphia, pp. 1-13.

Mislevy, R.J. (2008), "Some implications of expertise research for educational assessment", 34th International Association for Educational Assessment (IAEA) Conference, Cambridge, UK 8 September, 2008.

Mislevy, R.J. (1993), „Foundations of a new test theory", in N. Frederiksen, R. J. Mislevy and I.I. Bejar (eds.), *Test Theory for a New Generation of Tests*, Laurence Erlbaum, Hillsdale NJ, pp. 19-40.

Mislevy, R.J. et al. (2006) "Concepts, terminology, and basic models of evidence-centred design", in D.M. Williamson, R.J. Mislevy and I.I. Bejar (eds.), *Automated Scoring of Complex Tasks in Computer-Based Testing*, Laurence Erlbaum, Mahwah NJ, pp. 15-47.

Mislevy, R.J. et al. (1999), "A cognitive task analysis with implications for designing simulation-based performance assessment", *Computers in Human Behavior*, Vol. 15/3-4, pp. 335-374.

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

Paek, P.L. (2002), "Problem solving strategies and metacognitive skills on SAT mathematics items. (Doctoral dissertation, University of California, Berkeley, 2002)" *Dissertation Abstracts International*, Vol. 63/(9, 3139.

Pelligrino, J., N. Chudowsky and R. Glaser (eds.) (2001), *Knowing What Students Know: The Science and Design of Educational Assessment*, Committee on the Foundations of Assessment, National Research Council, National Academy Press, Washington, DC.

Phifer, S.J. and J.A. Glover (1982), "Don't take students' word for what they do while reading", *Bulletin of the Psychonomic Society*, Vol. 19., pp. 194-196.

Polya, G. (1945), *How to Solve It*, Princeton University Press, Princeton, NJ.

Polya, G. (1957), *How to Solve It*, 2nd edition, Princeton University Press, Princeton, NJ.

Quellmalz, E. and J.W. Pellegrino (2009), "Technology and testing", *Science*, Vol. 323/5910, pp. 75-79.

Richardson M., et al. (2002), "Challenging minds? Students' perceptions of computer-based World Class Tests of problem solving", *Computers in Human Behavior*, Vol. 18/6, pp. 633-649.

Rupp, A. (2002), "Feature selection for choosing and assembling measurement models: A building-block-based organisation", *International Journal of Testing*, Vol. 2/3-4, pp. 311-360.

Sternberg, R.J. and T. Ben-Zeev (2001), *Complex Cognition: The Psychology of Human Thought*, Oxford University Press, New York, NY.

Stevens, R.H. and V. Thadani (2006), "A Bayesian network approach for modeling the influence of contextual variables on scientific problem solving", *International Conference on Intelligent Tutoring Systems*, Springer, Berlin and Heidelberg, pp. 71-84.

Thadani, V., R.H. Stevens and A. Tao (2009), "Measuring complex features of science instruction: Developing tools to investigate the link between teaching and learning", *The Journal of the Learning Sciences*, Vol. 18/2, pp. 285-322.

Van Gog, T., F. Paas, J.J.G. Van Merriënboer and P. Witte (2005), "Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting" *Journal of Experimental Psychology: Applied*, Vol. 11, pp. 237-244.

Vendlinski, T. and R. Stevens (2002), "Assessing student problem-solving skills with complex computer-based tasks", *Journal of Technology, Learning, and Assessment*, Vol. 1/3, http://ejournals.bc.edu/ojs/index.php/jtla/article/viewFile/1669/1511, accessed 14 December 2012.

Wainer, H. and G.L. Kiely (1987), "Item clusters and computerized adaptive testing: A case for testlets", *Journal of Educational Measurement*, Vol. 24/3, pp. 185-201.

Williamson, D.M. et al. (2004), "Design rationale for a complex performance assessment", *International Journal of Testing*, Vol. 4/4, pp. 303-332.

Williamson, D.M., I.I. Bejar and R.J. Mislevy (2006), "Automated scoring of complex tasks in computer-based testing: An introduction", in D.M. Williamson, I.I. Bejar and R.J. Mislevy (eds.), *Automated Scoring of Complex Tasks in Computer-Based Testing*, Laurence Erlbaum, Hillsdale, NJ, pp. 1-13.

Wilson, M. and R.J. Adams (1995), "Rasch models for item bundles", *Psychometrika,* Vol. 60/2, pp. 181-198.

Wirth, J. and E. Klieme (2003), "Computer-based assessment of problem solving competence", *Assessment in Education*, Vol. 10/3, pp. 329-345.

Zoanetti, N.P. (2010), "Interactive computer based assessment tasks: How problem-solving process data can inform instruction", Australasian Journal of Educational Technology, Vol. 26/5, pp. 585-606, http://www.ascilite.org.au/ajet/ajet26/zoanetti.pdf,accessed 14 December 2012.

Chapter 12

Educational process mining: New possibilities for understanding students' problem-solving skills

By

Krisztina Tóth

Center for Research on Learning and Instruction, Szeged University, Hungary.

Heiko Rölke, Frank Goldhammer

German Institute for International Educational Research, Frankfurt, Germany.

and Ingo Barkow

Swiss Institute for Information Science (SII), University of Applied Sciences HTW Chur, Switzerland.

The assessment of problem-solving skills heavily relies on computer-based assessment (CBA). In CBA, all student interactions with the assessment system are automatically stored. Using the accumulated data, the individual test-taking processes can be reproduced at any time. Going one step further, recorded processes can even be used to extend the problem-solving assessment itself: the test-taking process-related data gives us the opportunity to 1) examine human-computer interactions via traces left in the log file; 2) map students' response processes to find distinguishable problem-solving strategies; and 3) discover relationships between students' activities and task performance. This chapter describes how to extract process-related information from event logs, how to use these data in problem-solving assessments and describes methods which help discover novel, useful information based on individual problem-solving behaviour.

The NEPS data collection is part of the Framework Programme for the Promotion of Empirical Educational Research, funded by the German Federal Ministry of Education and Research, and supported by the Federal States.

Introduction

In educational measurement the concept of problem-solving skills is of particular importance. They represent cross-curricular skills that are assumed to be involved in various specific domains like reading, mathematics and science, enabling one to master demands in real-life adult settings (compare Wirth and Klieme, 2003). Accordingly, the assessment of problem-solving skills has become an important goal of the Organisation for Economic Co-operation and Development's (OECD) Programme for International Student Assessment (PISA) and its Programme for the International Assessment of Adult Competencies (PIAAC).

Problem solving is not viewed as a one-dimensional construct, but can conceptually be decomposed into analytical aspects and dynamic or interactive aspects. For instance, PISA 2003 assessed students' analytical problem-solving skills using static problems. Students were required to complete paper-based tasks where all the information needed to obtain a solution was complete and transparent to the problem solver (compare OECD, 2004). In PISA 2012, however, there was a move to assessing interactive problem-solving skills, requiring students to deal with problems where they had to uncover information by interactively exploring the problem situation (compare OECD, 2010).

The problem-solving construct PIAAC assessed refers to "information-rich" problems requiring solvers to locate and evaluate information by interacting with simulated software applications, such as a web browser (OECD, 2012). Constructing valid interactive problem-solving measures requires a computer-based assessment strategy containing complex and interactive tasks, as the problem solver needs to explore and perhaps manipulate a (simulated) problem situation. At the behavioural level, such measures do not only provide product data (i.e. the correct or incorrect response), but also rich process data about the test taker's path to finding a solution. One promising analysis strategy for exploiting this huge amount of unstructured data is called process mining.

Process mining enables scientists to discover new forms of data analysis in educational assessment. While classical measurement (e.g. paper-and-pencil tests) can only provide insights from analysing the end results, process mining can trace students' abilities to solve the tasks. This provides a more granular picture of the students' skills, as the way they solve the problems supplies valuable information above and beyond the results. Therefore, process mining as a method can open new ways of analysing the students' problem-solving behaviour. Nevertheless, the implementation of process mining can be considered an interdisciplinary challenge requiring subject matter experts, item developers, psychometricians and computer scientists to work together to extract, aggregate, model and interpret the data appropriately. The following sections discuss the first steps and approaches involved.

This chapter is divided into five parts. First it gives a brief overview of the type of information that can be extracted from tracked data, and describes how to extract useful information from logs. The next section presents the process-mining method that allows the analysis of process data retrieved from different educational settings. These background sections are followed by defining the research objectives that can typically be addressed by means of process mining. In accordance with the main objective of this chapter, the next section provides insights into applying process mining to problem-solving behaviour data (e.g. visualisation, clustering, and classification). Finally it describes the implications for online problem-solving assessments.

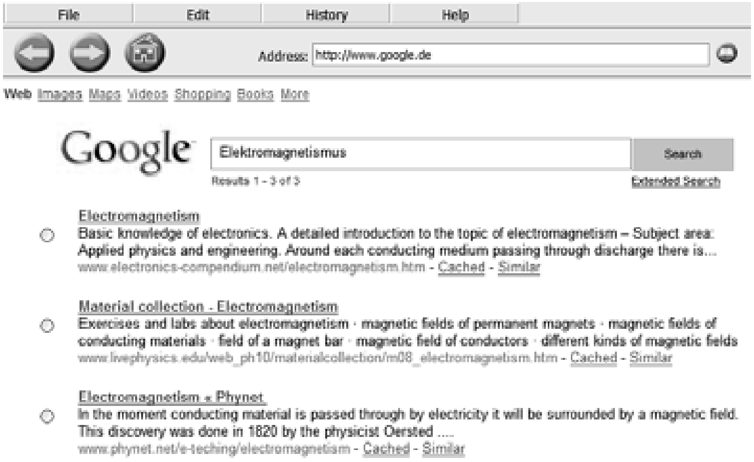
Background: From logs to knowledge

The way people use their knowledge during assessment cannot be directly observed but the outcomes can be derived from task results. Moreover, students' actions in a computerised environment can be automatically traced as an indicator of the process of using knowledge (Chung and Baker, 2003). These tracking data (logs) support the observation of students' online behaviour and enable researchers to examine problem-solving processes and other phenomena.

These log data are automatically accumulated during the assessment, as all the actions of the students are usually stored in generated log files or databases. The process of recording students' interactions (e.g. clicks, key presses and drags/drops) with a test delivery platform is called data logging. It allows one to reproduce and investigate students' individual online activity (for example changing an answer, using an embedded link or moving a slider).

The test item in Figure 12.1 demonstrates a problem situation designed by Pfaff and Goldhammer (2012) simulating a web search engine. The opening screen of this interactive task displays a list of webpages according to the query (using the keyword "electromagnetism") with short summaries containing documents' titles and small parts of the hypertext. Three connected websites can be accessed from the list of search results in Figure 12.1 via hyperlinks.

Figure 12.1. **Web search item**



○ Electromagnetism
Basic knowledge of electronics. A detailed introduction to the topic of electromagnetism – Subject area: Applied physics and engineering. Around each conducting medium passing through discharge there is...
www.electronics-compendium.net/electromagnetism.htm - Cached - Similar

○ Material collection - Electromagnetism
Exercises and labs about electromagnetism · magnetic fields of permanent magnets · magnetic fields of conducting materials · field of a magnet bar · magnetic field of conductors · different kinds of magnetic fields
www.livephysics.edu/web_ph10/materialcollection/m08_electromagnetism.htm - Cached - Similar

○ Electromagnetism « Phynet
In the moment conducting material is passed through by electricity it will be surrounded by a magnetic field. This discovery was done in 1820 by the physicist Oersted ....
www.phynet.net/e-teching/electromagnetism - Cached - Similar

Source: Pfaff and Goldhammer (2012), "Evaluating online onformation: The influence of cognitive components".

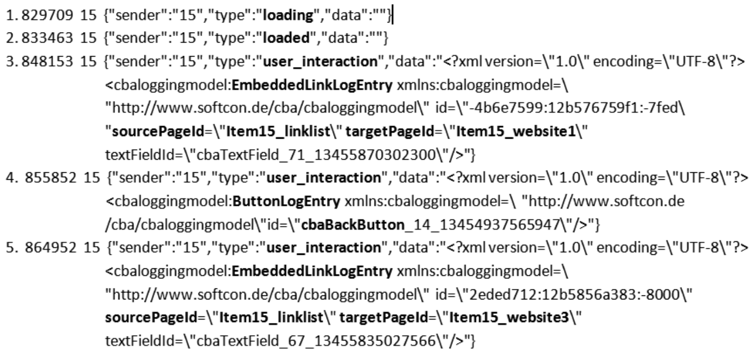
Figure 12.2 depicts possible human-computer interactions in this simulation in the form of part of a log file. It provides a student's actions in an extended markup language (XML) schema and it is composed of five log entries (lines). The first digit in each line is a line reference. The second number is a time stamp (in milliseconds) giving the time the event happened. The third number ("15") is the task id. The character stream on the fourth position in each line describes an action during the assessment process.

The first line of Figure 12.2 shows the beginning of Task 15 (loading data). The second line shows the time at which all the information of the hypertext item appeared to the student (loaded data). The third records a page request, when the student asked to load the Item15_website1 after 848 153 milliseconds. This student had explored the opening screen of the task containing the lists of the web search (source page: Item15_linklist) and clicked on the first embedded link available in the search results list. After exploring this webpage, the student navigated back to the list of links with a back button (line 4: cbaBackButton). This was followed by visiting the third website (Item15_website3) depicted in line 5.

The results of related works on investigating students' learning or test-taking behaviour in online environments (for example, see Hershkovitz and Nachmias, 2011, Mazza and Milani, 2005;

Pechenizkiy et al., 2009) are based on similar log information accumulated during online assessment. But these raise the following questions (see Tóth, Rölke and Goldhammer, 2012a): How can the logs be analysed? How can educational researchers gain insights or at least useful information from students' activities extracted from the individual assessment processes?

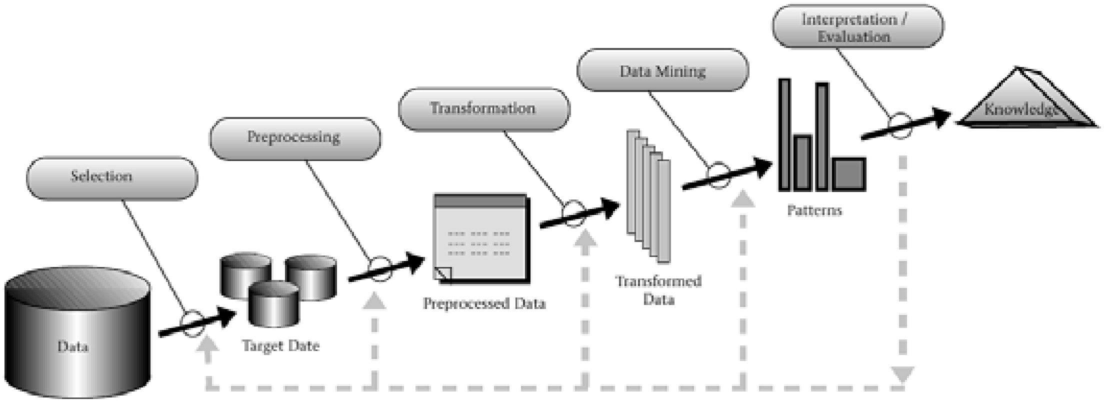
Figure 12.2. **Sample log file**



```
1. 829709 15 {"sender":"15","type":"loading","data":""}
2. 833463 15 {"sender":"15","type":"loaded","data":""}
3. 848153 15 {"sender":"15","type":"user_interaction","data":"<?xml version=\"1.0\" encoding=\"UTF-8\"?>
             <cbaloggingmodel:EmbeddedLinkLogEntry xmlns:cbaloggingmodel=\
             "http://www.softcon.de/cba/cbaloggingmodel\" id=\"-4b6e7599:12b576759f1:-7fed\
             "sourcePageId=\"Item15_linklist\" targetPageId=\"Item15_website1\"
             textFieldId=\"cbaTextField_71_13455870302300\"/>"}
4. 855852 15 {"sender":"15","type":"user_interaction","data":"<?xml version=\"1.0\" encoding=\"UTF-8\"?>
             <cbaloggingmodel:ButtonLogEntry xmlns:cbaloggingmodel=\ "http://www.softcon.de
             /cba/cbaloggingmodel\"id=\"cbaBackButton_14_13454937565947\"/>"}
5. 864952 15 {"sender":"15","type":"user_interaction","data":"<?xml version=\"1.0\" encoding=\"UTF-8\"?>
             <cbaloggingmodel:EmbeddedLinkLogEntry xmlns:cbaloggingmodel=\
             "http://www.softcon.de/cba/cbaloggingmodel\" id=\"2eded712:12b5856a383:-8000\"
             sourcePageId=\"Item15_linklist\" targetPageId=\"Item15_website3\"
             textFieldId=\"cbaTextField_67_13455835027566\"/>"}
```

Tóth et al. (2012a:2) address these objections by arguing that the process of extracting useful information from large datasets is consistent with the knowledge discovery process in databases of Fayyad et al. (1996) presented in Figure 12.3. First, one has to select the relevant pieces of information that may help to answer research questions from the large amount of log entries (Fayyad et al., 1996). In this phase, information is extracted from the logs and a target dataset is created.

Figure 12.3. **The knowledge discovery process**



Source: Fayyad et al. (1996), "From data mining to knowledge discovery: An overview".

Afterwards, the target data are preprocessed, i.e. cleaned (for example handling missing data and removing noise) and transformed (normalised and aggregated) into the format required by statistical analysis software (Romero, Ventura and García, 2008). This transformation step is followed by the analysis phase, choosing the most appropriate analysis methods and applying them to obtain patterns, models or results. The data analyses can be accomplished by applying data-mining techniques (see Baker and Yacef, 2009), taking the research aims and the target dataset into account. The final step is the interpretation and evaluation of (educational) data (Fayyad et al., 1996). As the arrow in this model shows, the knowledge-discovery process is iterative, allowing one to return to previous steps for more effective knowledge discovery.

Analysing problem-solving behaviour data: Educational process mining

The term process mining refers to "the development of a set of intelligent tools and techniques aimed at extracting process-related knowledge from event logs recorded by an information system" (Pechenizkiy et al., 2009:1). It makes it possible to extend the examination of dichotomous test results, which simply reflect whether the student solved the task correctly or not, with online behaviour data. Furthermore, it "can be seen as a sub-domain of data mining" (van der Aalst and Weijters, 2004:239) which requires data that are (totally) ordered in predefined way. The ordering may be given explicitly in terms of time or implicitly in terms of a sequence. Exploiting this ordering, process-mining algorithms try to (re-)construct models based on observed interaction paths within the assessment system.

If process mining techniques are applied to educational data, they are often referred to as educational process mining (EPM; for example by Tröka and Pechenizkiy, 2009). There are three types of EPM: 1) process model discovery; 2) conformance analysis; and 3) process model extension (Pechenizkiy et al., 2009).

Process model discovery seek to construct a model based on log events (van der Aalst, 2011) which describes the log data in a compact and concise way. Process model discovery therefore often produces formal models using graphical representations like finite state machines or Petri nets. Using these visual representations, helps item and test developers to better understand how test takers actually interacted with their items and where to look for possible item functioning issues.

Conformance analysis requires an existing theoretical model that can be compared with the discovered model built from log events (van der Aalst, 2011). It can therefore confirm whether the observed behaviour is in accordance with the model assumptions (Pechenizkiy et al., 2009). For instance, in the case of a complex problem-solving item, there may be several ways of solving the problem, depending on the level of expertise of the test taker. The question is whether these ways are reflected in the log data as expected. It might be the case that certain solutions are never tried or the other way round, that additional solutions are possible which the item developer did not foresee.

Process model extension also involves an existing theoretical model. But in contrast with conformance analysis, in which the expectations are compared with the tracked data, process model extension (called enhancement) seeks to extend or improve process models based on observed behaviour data (van der Aalst, 2011).

While the roadmap for applying all three types of process mining to problem-solving assessments is clearly laid out, its implementation is still a work in progress. To provide an overview of the possibilities of educational process mining this chapter focuses on the first type of process mining, which is process model discovery.

Objectives

The main aim of this chapter is to show how test-taking process data can be integrated into educational assessment. What follows attempts to show how process data can be used in problem-solving assessments to 1) examine human-computer interaction via traces left in the log file; 2) map and cluster students based on response processes to find different problem-solving behaviour patterns; and 3) discover relationships between test takers' activities and their test performance.

Applying process mining to problem-solving behaviour data

Process mining methods include visualisation, clustering and classification. This section briefly describes these methods and their application in learning and/or assessment environments, including related research work, and discusses their use in problem-solving assessment.

Visualisation: Examining human-computer interactions

With large datasets such as the test-taking process data generated by complex test instruments, getting an overview of the students' observed activities is an important first step before data analysis starts. Information visualisations producing graphical representations from the data may be an appropriate way to do this (Romero and Ventura, 2010). A variety of visualisation techniques can generate such graphical depictions (Keim, 2002) and serve the data in more easily comprehensible form (Mazza and Dimitrova, 2003), give an overview of students' real test-taking behaviour (Ferreira de Oliveira and Levkowitz, 2003) and support researchers in data exploration. The findings of the visualisation can then then be used to choose the best methods to analyse the data, from which new hypotheses will eventually emerge (Keim, 2002).

Data visualisation has mostly been exploited in distance learning (Mazza and Milani, 2005) or course management systems (Mazza, 2006; Zhang et al., 2007). But it is rarely used to mine process-related educational assessment data. For this reason, we attempt here to provide insights into the representations that might be used in educational assessment, especially in problem-solving assessment.

Two visual representations of a "web browser" item (Figures 12.4 and 12.5) can highlight the power of visualisations in problem solving assessment, The visualisations are taken from a study of Tóth et al. (2012a:4-5) based on items developed by Pfaff and Goldhammer (2012) and administered to students in the National Educational Panel Study (NEPS), similar to the item presented in Figure 12.1. These items simulate web search engines within various contexts (e.g. health, education or sports); the problem solver must find the most relevant web page. The start screen of each item displays a list of search results with short summaries (hits), and for each item, five connected websites are accessible via hyperlinks.

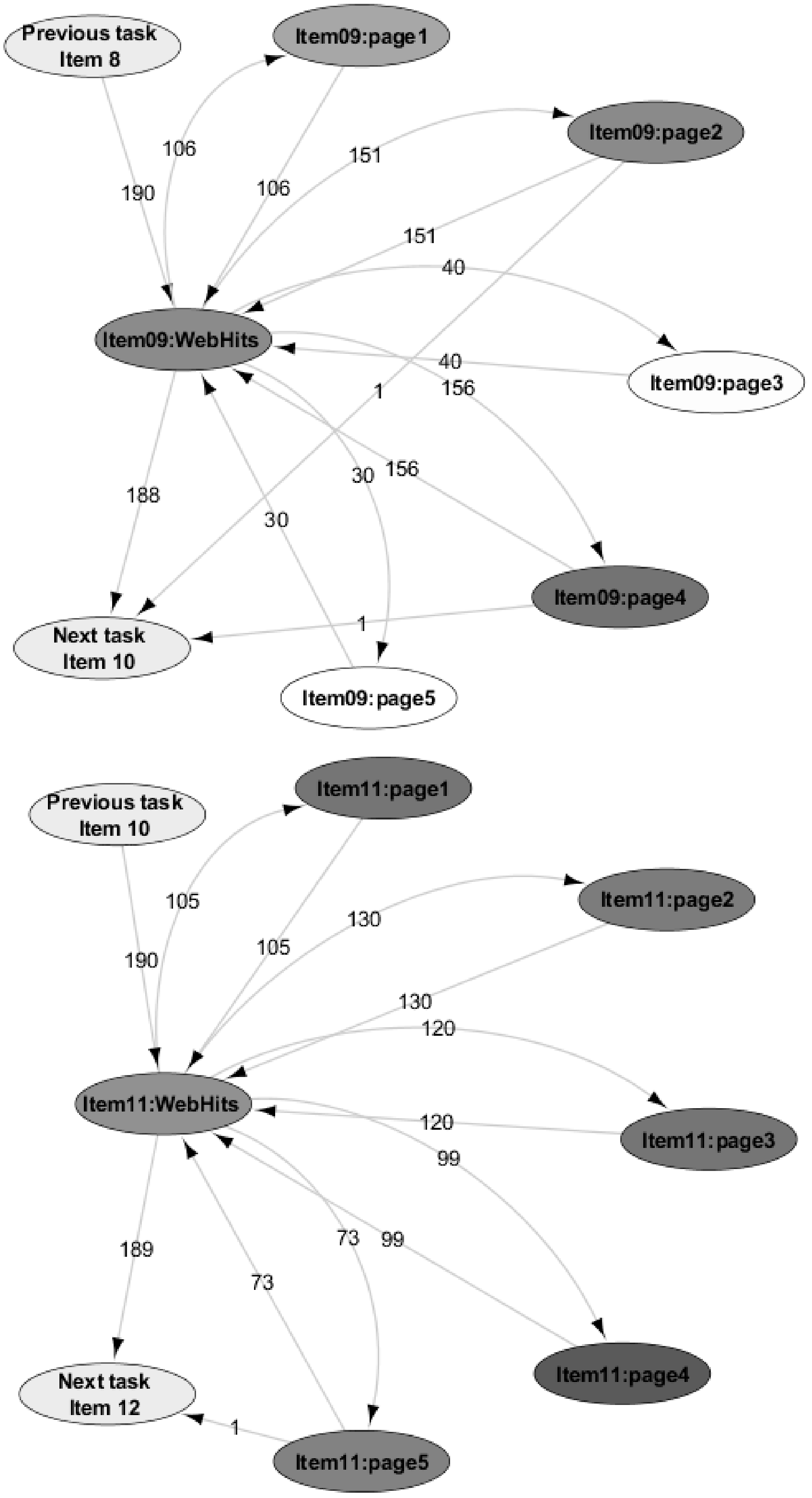
To give an overview of test takers' problem-solving activity at the item level we aggregated the observed problem-solving processes. In Figure 12.4, test takers' problem-solving activities are depicted as directed graphs. The nodes of the graphs represent the web search result page (WebHits) with the available hyperlinks and websites (page1, page2 etc., in Figure 12.4), connected to WebHits.

The nodes are connected via directed edges. The edge weights show the frequency of page visits originating from the source page. All items began with the list of links and the task definition page. For example, on Item 9, the opening screen presents the results list of a web search (Item9:WebHits node). On this website, test takers most frequently clicked on the fourth embedded link available in the web search results' list, and explored the fourth webpage (Item9:page4). All of the page 4 visits were followed by navigating back to the web hits page. The second most frequently visited page is page 2, and so forth. The most rarely explored page is the last (Item09:page5).

The graphical representation of the navigation-related information was extended by time-related information, i.e., by the time spent on various webpages. In Figure 12.4, the node colour refers to the average time spent on the website: the nodes are shaded from dark green to white (previous and subsequent tasks are coloured yellow because they play no role in this study). White indicates a short time spent on a page while dark green represents a long time. In the case of Item 9, test takers spent the least time on the third and fifth pages and the most time on the fourth page. For Item 11, test takers did not show any significant difference on the average time spent exploring the different websites.

The graph representing test takers' navigation behaviour on Item 9 thus shows that test takers explored the third and fifth website less often and spent less time on these sites compared to other webpages. In contrast, the graph of Item 11 reveals a balanced distribution of page visits. Presumably, test takers perceived the alternative links (and the short summaries on the web search results' page) as fairly equally likely to represent the best answer, as indicated by the frequency of page visits.

Figure 12.4. **Aggregated test-taking processes of students on Item 9 and Item 11 (directed graph)**



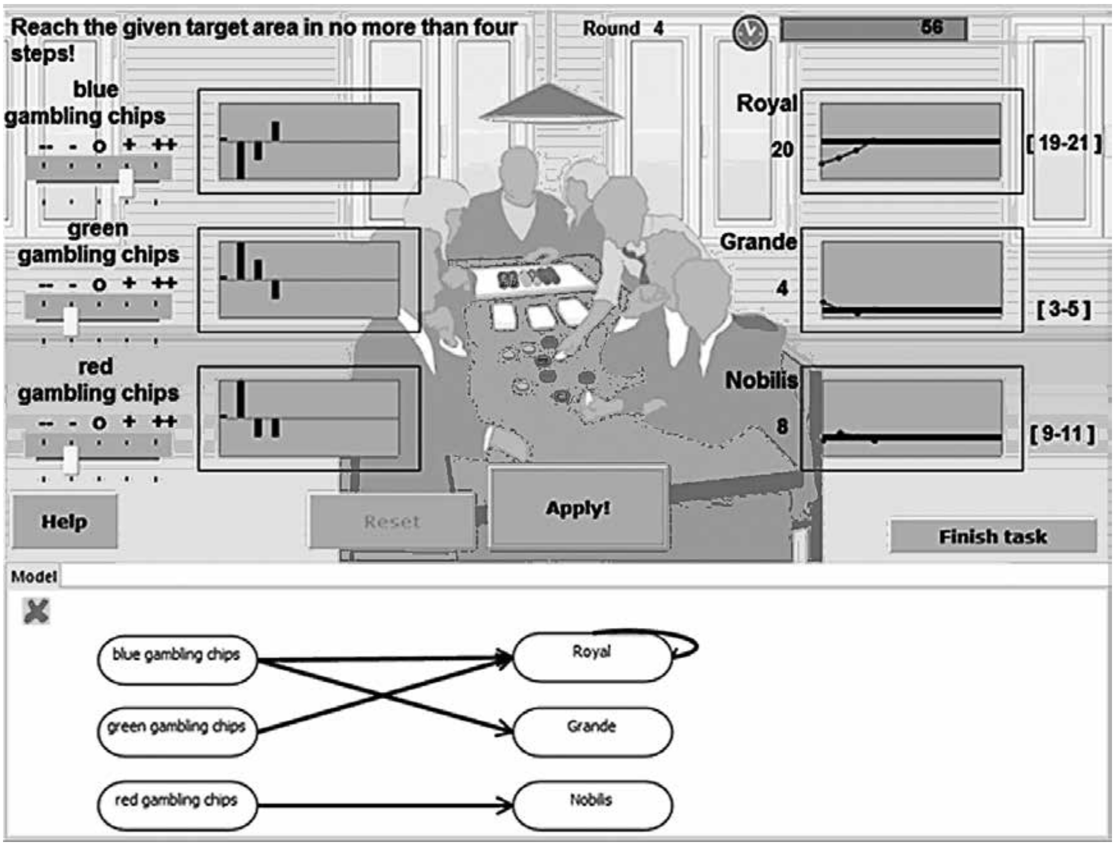
Source: Tóth et al. (2012a), "Investigating test-taking behaviour in simulation-based assessment: Visual data exploration".

To get a more detailed picture of human-computer interactions at the individual level in the web browser environment, we plotted test takers' activities according to the time series with the dotted chart analyser of the ProM framework (van Dongen et al., 2005). This tool allows all of the individual link selection activities to be plotted with time information (for the details see Tóth et al., 2012a).

Other types of problem-solving tasks aim to measure other aspects of problem solving (for example, see Greiff and Funke, Chapter 6; McCrae et al., Chapter 14; OECD, 2012, 2014) and require other kinds of visual representation. To illustrate this, consider students' problem-solving behaviour in tasks using the MicroDYN approach (Greiff and Funke, 2009). These tasks measure three dimensions of complex problem solving: exploration, model building, and forecasting (Tóth et al., 2012c). First students explore the system; then they depict the connections between variables; and finally try to reach the predefined values of the target variables.

The item presented in Figure 12.5 is an example of the MicroDYN approach. In this simulation test takers have to find relationships between gambling chips and play of cards (see Greiff et al., 2013). To test the connection between exogenous variables (the amount of blue, green and red gambling chips) and the endogenous ones (different kinds of playing cards labeled Royal, Grande and Nobilis) test-takers drag the sliders to decrease and increase the input variables. After moving the sliders, test-takers click the "Apply" button to execute the manipulation of the situation. Pressing the apply button defines one execution (also called a round) in the assessment. In parallel with their exploration of the system, students mark the relationship between exogenous and endogenous variables (model building), then try to reach the given area of the target variables (forecasting).

Figure 12.5. **Example of a MicroDYN item**

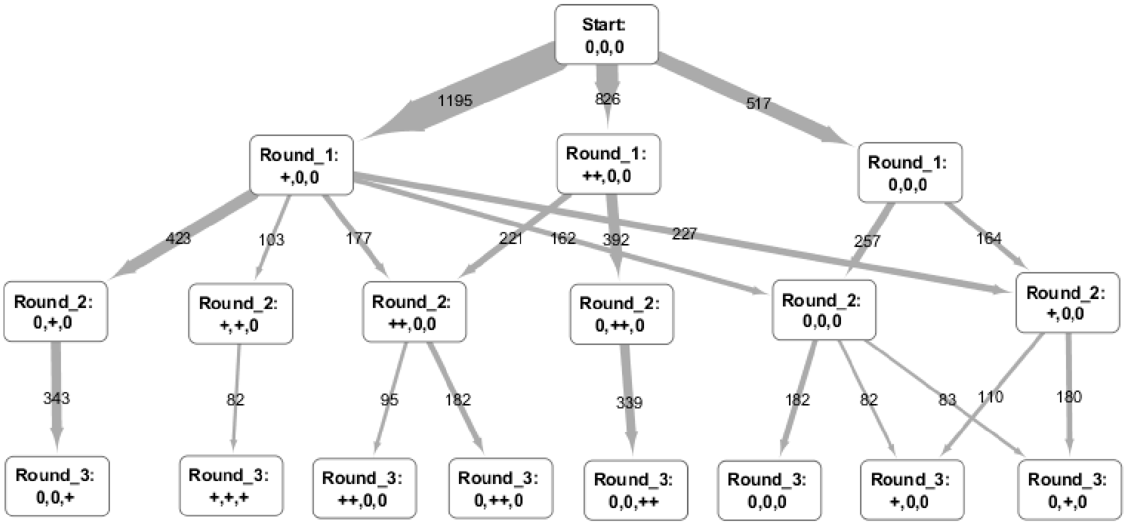

Reach the given target area in no more than four steps!

Source: Greiff, S., et al. (2013), "Complex problem solving in educational settings – Something beyond g: Concept, assessment, measurement invariance, and construct validity", *Journal of Educational Psychology*, Vol. 105/2, pp. 364-379, http://dx.doi.org/10.1037/a0031856.

Figure 12.6 visualises the solution-path behaviour as a directed graph (Tóth, Greiff, Wüstenberg and Rölke, 2014), which shows the activity of test takers in a test composed of seven problem-solving items. First, the problem-solving activities were aggregated across the seven items and across all test takers. Then the accumulated problem-solving processes of the first three executions were represented in a tree structure. The tree is composed of nodes and directed edges. The node on the top of the tree is called the root node, and represents the start of the task. The edges show the process of the problem solving. Lastly, the weights of the edges present the frequency of the same execution (only edges with values greater than 80 are shown in the graph). The initial value of the three slider handles was 0 (depicted as Start: 0,0,0).

Test takers' actions in the first execution (Round_1) are presented in the first depth of the tree. The actions are presented with three signs delimited by commas, representing the three slider positions. For example "Round_1: +,0,0" represents a test takers moving the first slider to +, and leaving the second and third sliders at their starting position.

Figure 12.6. **Aggregated problem solving behaviour on seven tasks, activities in the first three executions**



Clustering students based on problem-solving behaviour

The objective of cluster analysis or clustering is to group similar objects into classes called clusters (Romero and Ventura, 2010). In educational research, clustering is used to discover similarly behaving groups of students interacting with an e-environment. For instance, Talavera and Gaudioso (2004) used cluster analysis to identify student profiles based on online behaviour data, and Perera and colleages (2009) clustered learners and teams in a collaborative environment in order to discover subgroups of students requiring different interventions. Cobo and colleagues (2011) proposed finding different behaviour patterns in online discussion forums.

Clustering human behaviour is often based on previously defined process measures (as in Talavera and Gaudioso, 2004; Perera et al., 2009). These process measures are derived from individual test-taking behaviour and seek to characterise test-takers' actions in online environments. These measures (called features or observables by Bennett et al., 2010) can be used to compare students' online behaviour. The advantage of applying predefined process measures is that they facilitate the interpretation of results (Mant, 2001). For this reason, we performed a cluster analysis with predefined features in problem-solving behaviour analysis (see for example Tóth, Rölke, Greiff and Wüstenberg, 2014).

In the problem-solving assessment described above, we selected a web-browser item with five available links for the cluster analysis to discover similarly behaving subgroups of test takers. Table 12.1 lists the set of process measures used to cluster the students.

Table 12.1. **Process measures to cluster students**

Process measures	Interpretation
1. Number of page visits	How many web sites visits were performed during the test session, including new page visits and revisits.
2. Number of different page visits	How many different websites were accessed by embedded links from the web search result page.
3. Visit to relevant page	Whether the page which contained the relevant pieces of information required to complete the task correctly was visited or not (yes/no).
4. Completion time	Time spent on an item
5. Ratio of time spent on the relevant page	Time spent on the relevant page divided by completion time
6. Ratio of time spent on the opening screen	Time spent on the web search result page divided by the completion time

Source: Tóth, Rölke, Goldhammer and Naumann (2012b), "Investigating problem solving behaviour in simulated hypertext environments".

To investigate similar problem-solving behaviour we used the K-means clustering algorithm with Euclidean distance function in the WEKA data mining software package (Hall et al., 2009). This algorithm identified three clusters of test takers (see Table 12.2).

Most test takers in the first subgroup (Cluster 1) failed to find the correct response. Only 19.15% of the members of Cluster 1 solved the task correctly. These test takers did not visit the relevant website during the test session, only used a small number of hyperlinks and spent little time on tasks. So they probably tried to answer the questions based on information about hits like the domain name and the short summary of the websites.

Table 12.2. **Clustering test takers based on problem-solving behaviour in an online environment**

Process measures	Full data	Cluster centroids		
		Cluster 1	Cluster 2	Cluster 3
Number of page visits	5.18	2.19	8.49	4.30
Number of different page visits	2.38	0.94	3.84	2.08
Visit to relevant page (Yes/No)	Yes	No	Yes	Yes
Completion time	67.74	42.16	93.06	62.76
Time on the relevant page	0.14	0	0.18	0.19
Time spent on the opening screen	0.64	0.83	0.49	0.66
Ratio of correct responses	76.19%	19.15%	96.83%	93.67%
Clustered instances	–	0.25	0.33	0.42

The probability of correct responses is high among test takers in Cluster 2 (96.8%) and Cluster 3 (93.7%). However, test takers in Cluster 2 needed more time to find the relevant website and they used more embedded links and explored more webpages than those in Cluster 3, who looked at only a few websites and spent more time on the hit page than Cluster 2 test-takers. So the problem-solving behaviour patterns of the last cluster identify test takers with better problem-solving skills, as they needed less time and fewer page visits to find the information required in the task than those in Cluster 2, without significantly decreasing their probability of success.

The process measures defined for this process-centric analysis of problem solving behaviour are rarely transferable to other types of items. Because of the specific task characteristics, new task-specific, measures have to be defined for each type. For instance, Table 12.3 lists the features which appear to be relevant for MicroDYN items.

Table 12.3. **Process measures describing test takers' activities for MicroDYN items**

Process measures	Interpretations
1. Number of executions	The frequency of using the apply button.
2. Ratio of repeated executions	Number of executions which were used previously on the item, divided by the total number of executions.
3. Ratio of rounds with one exogenous variable	Number of executions in which one exogenous variable was manipulated, divided by the total number of executions,
4. Exploration time	Time given in milliseconds spent in the exploration phase.
5. Ratio of zero rounds	The frequency of executions in which all sliders stand at 0, divided by the number of execution.

Source: Tóth et al. (2012c), "Prediction of students' performance on test taking processes in Complex Problem Solving".

Based on these predefined process measures (except for the ratio of zero rounds, which is relevant only for classification) we sought to split the group of test takers who had created the correct mental model on the item presented in Figure 12.5 into more and less efficient subgroups. Table 12.4 lists the results of the K-means clustering according to Tóth et al. (2014). The more efficient test takers included in Cluster 1 solved the task in 66.10 seconds, while those in Cluster 2 needed more time to find the correct mental model; Cluster 1 members also needed fewer executions than Cluster 2 members. Test takers in Cluster 1 avoided many repeated executions (only 24% of rounds), as opposed to Cluster 2 ones who reconsidered the same exogenous variable quite often (71%).

Table 12.4. **Clustering students based on online behaviour on the MicroDYN item**

Process measures	Full data	Cluster centroids	
		Cluster 1	**Cluster 2**
Number of rounds	14.3	6.54	20.65
Ratio of repeated executions	0.5	0.24	0.71
Ratio of rounds with one exogenous variable	0.55	0.74	0.39
Exploration time	74.31	66.09	81.03
Clustered instances	–	0.45	0.55

Source: Tóth et al. (2012c), "Prediction of students' performance on test taking processes in Complex Problem Solving".

Cluster analysis using predefined process measures thus identifies similarly behaving subgroups of student. In the first example, test takers could be grouped into three clusters (more or less efficient and inefficient). The second example showed how efficient test takers could be further partitioned.

Classification: Relationship between problem-solving behaviour and test performance

Classification is the technique of placing objects into predefined groups or classes based on their attributes (Hämäläinen and Vinni, 2010). This is one of the most valuable techniques used in educational process mining (Romero, Ventura, Espejo and Hervás, 2008). It can be used to classify the performance of test takers (Zoanetti, 2010), predict success (Romero et al., 2010) or detect disengagement in an online environment (Cocea and Weibelzahl, 2009).

There are several classification approaches available. These include decision trees, neural networks, rule induction and fuzzy rule induction (see Hämäläinen and Vinni, 2010; Romero, Ventura and García, 2008). This chapter looks at using decision trees to classify test takers' problem-solving behaviour on a PIAAC test item (see Figure 12.7) that required test takers to acquire and evaluate information by interacting with a simulated web search engine. The task was to select the website(s) which matched the set of predefined search criteria (OECD, 2012). The opening screen describes the task on the left and the results of the web search on the right. From this page test takers had to navigate through the hypertext documents connected via hyperlinks to locate the appropriate website.

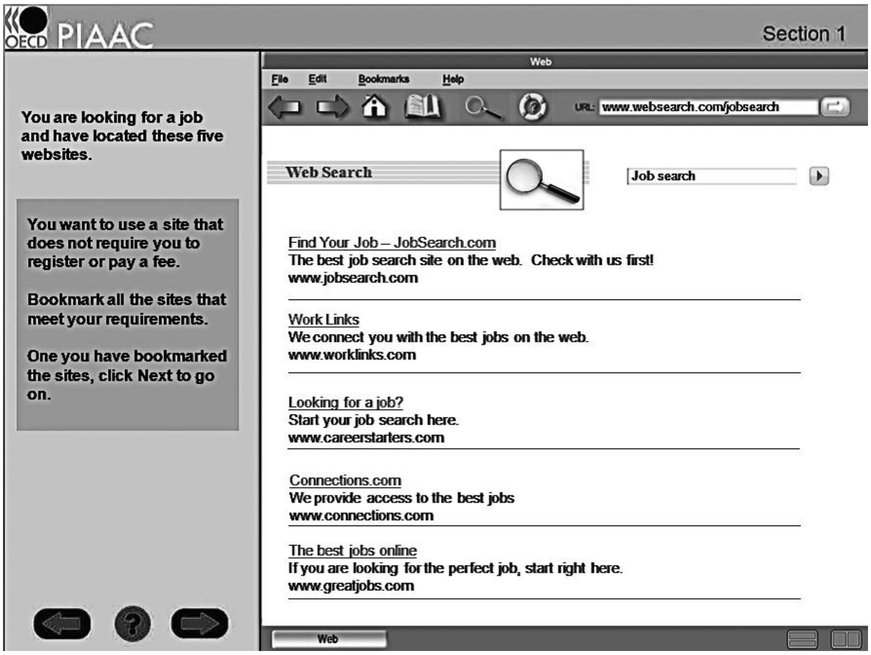
For the analysis, similar process measures were defined as were used earlier to cluster the actions of test takers on the web browser test in Table 12.2: number of page visits, number of different page visits, relevant page visits and the ratio of time spent on the relevant pages, with the addition of a variable measuring the number of bookmarked pages. The classification was performed using the C4.5 algorithm (see Quinlan, 1993) implemented as a J4.8 algorithm in WEKA. This decision-tree algorithm creates a tree-like model depicted in Figure 12.8 which visualises the relationship among different variables. The decision tree consists of decisions (depicted as oval nodes) and class labels (leaves). Each decision node evaluates an attribute (see predefined measures), the decision determines which path to follow. Leaf nodes stand for the classification of test performance: IC for incorrect and C for a correct response.

Clearly, all paths from from the root node to the different leaves define a rule set. Based on the analysis reported in Tóth et al. (2012b) the number of different pages visited (represented as the root node in Figure 12.8) appears to be the best predictor in this complex task. Test takers who visited fewer than eight different pages during the assessment had a high chance of getting an incorrect answer. Among this group, 111 test takers in the observed data made an incorrect response and only one of them was correct.

The specific activity required in the task (bookmarking) also offered high information gain. Test takers who visited more than eight different pages and bookmarked either more or fewer than two webpages did not make the correct response. This simple model in Figure 12.8 classifies 96.70% of the results correctly.

This classification method could be also adapted to other types of problem-solving tasks, but it requires process measures to describe the problem-solving behaviour in another item type in an effective way. Another example exploring the relationship between problem-solving (model building) performance and problem-solving behaviour, uses test-taking behaviour data on the MicroDYN item shown in Figure 12.5. For this analysis we applied the predefined process measures listed in Table 12.3: number of executions, ratio of repeated executions, ratio of rounds with one exogenous variable, exploration time and ratio of zero rounds.

Figure 12.7. **Job search task in PIAAC pre-test study**



You are looking for a job and have located these five websites.

You want to use a site that does not require you to register or pay a fee.

Bookmark all the sites that meet your requirements.

One you have bookmarked the sites, click Next to go on.

Find Your Job – JobSearch.com
The best job search site on the web. Check with us first!
www.jobsearch.com

Work Links
We connect you with the best jobs on the web.
www.worklinks.com

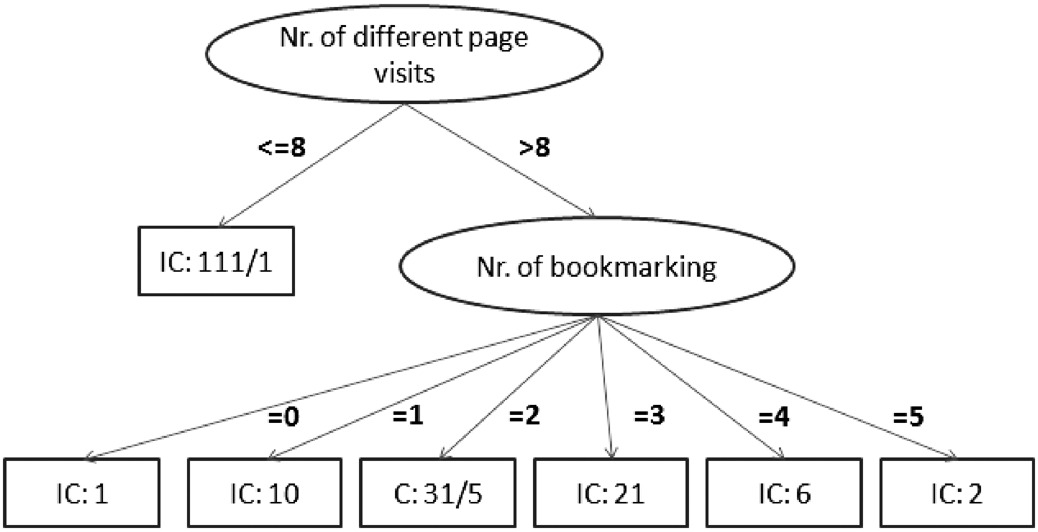
Looking for a job?
Start your job search here.
www.careerstarters.com

Connections.com
We provide access to the best jobs
www.connections.com

The best jobs online
If you are looking for the perfect job, start right here.
www.greatjobs.com

Source: OECD (2012), *Literacy, Numeracy and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills*, p. 53.

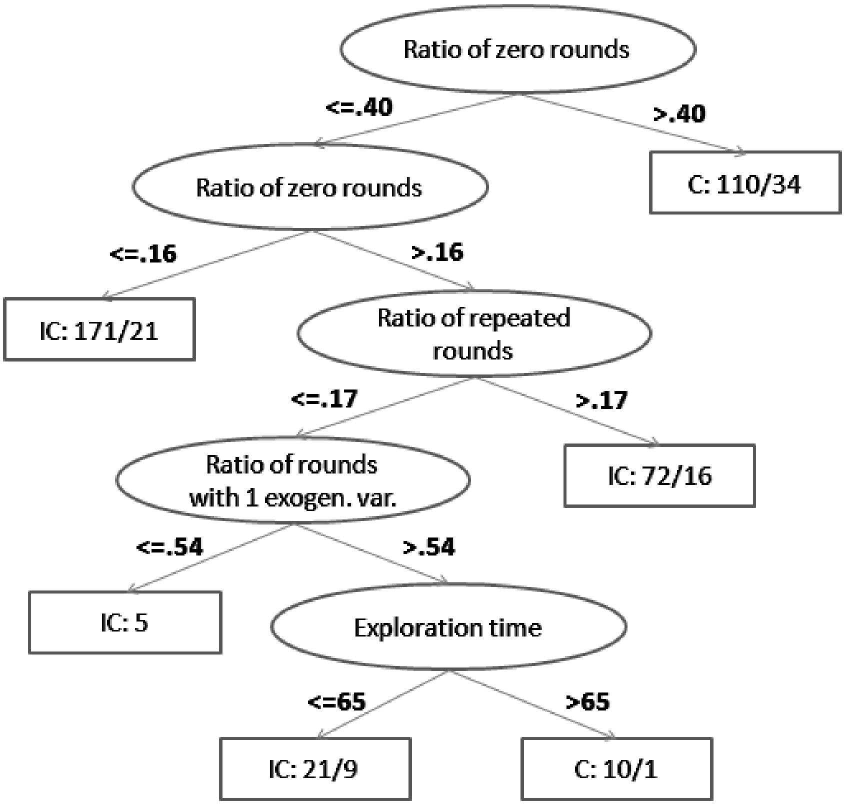
Figure 12.8. **Decision tree for the job search task**



Source: Tóth et al. (2012b), "Investigating problem solving behaviour in simulated hypertext environments".

Based on this feature set, Figure 12.9 shows the result of using the J48 algorithm according to Tóth et al. (2012c). In this task, the ratio of zero rounds gave the highest information gain, as the task contained eigendynamic elements (i.e. the Royal variable increased independently of any manipulations). Test takers using zero rounds for more than 40% of their executions in order to test the influence of any eigendynamic, presumably had a correct mental model whereas if the ratio of zero rounds was less than 16%, test takers probably had an incorrect mental model. The ratio of repeated executions, the ratio of rounds only one exogenous variable and exploration time variables were also useful for classification. The decision tree in Figure 12.9 classifies 79.18% of the test takers correctly.

Figure 12.9. **Decision tree for a complex problem-solving task**



Source: Tóth et al. (2012b), "Investigating problem solving behaviour in simulated hypertext environments".

Overall, classification with a decision tree gave insights into exploratory data analysis by placing students into classes based on predefined test-taking behaviour attributes. These data explorations have the advantage of depicting the relationship between problem-solving behaviour and test performance in an easily-interpreted tree structure and automatically identifying the most important process measures, i.e. those in the top nodes.

Implications for online problem-solving assessment

This study investigated using process-mining techniques in a problem-solving assessment context, providing insights into process-model discovery. It gave examples of 1) examining and visualising human-computer interactions based on log traces using directed graphs and dotted charts; 2) clustering students' response processes to find different patterns of problem-solving behaviour patterns; and 3) discovering relationships between activities and test outcomes using decision trees. Of course, there are other algorithms that can be used for clustering and classification, such as Bayesian networks, and analysis methods such as association rule mining, sequential mining, which could be adapted to process exploration of tracking data. Future studies comparing methods for analysing log data are therefore recommended.

Other analysis strategies use process mining techniques not just in an exploratory way, but also to confirm model assumptions and, if the assumptions are not confirmed, to refine the model (model extension). Conformance checking compares a pre-existing model with a discovered one built from log events. For example, in the case of cross validation, conformance checking can be used to compare student models taken from various data collections. For instance, a model created from a small student sample could be checked against a model based on a larger data collection. Such conformance checking enables researchers to evaluate how the results from one data collection can be generalised to another. Another possible application might be to compare test-taking paths, to determine whether the log data reflects the path expected by item developer.

References

Baker, R.S.J.D. and Y. Yacef (2009), "The state of educational data mining in 2009: A review and future visions", *Journal of Educational Data Mining*, Vol.1/1, pp. 3-17.

Bennett, R.E., H. Persky, A. Weiss and F. Jenkins (2010), "Measuring problem solving with technology: A demonstration study for NAEP", *Journal of Technology, Learning, and Assessment*, Vol. 8/8.

Chung, G.K.W.K. and E.L. Baker (2003), "An exploratory study to examine the feasibility of measuringproblem-solving processes using a click-through interface", *Journal of Technology, Learning, and Assessment*, Vol. 2/2.

Cobo, G., D. et al. (2011), "Modeling students' activity in online discussion forums: A strategy based on time series and agglomerative hierarchical clustering", in M. Pechenizkiy, et al. (eds.), *Proceedings of the 4th International Conference on Educational Data Mining,* Eindhoven, 6-8 July 2011, pp. 253-258.

Cocea, M. and S. Weibelzahl (2009), "Log file analysis for disengagement detection in e-Learning environments", *User Modeling and User-Adapted Interaction*, Vol. 51/4, pp. 341-385.

Ferreira de Oliveira, M.C. and H. Levkowitz (2003), "From visual data exploration to visual data mining: A survey", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 9/3, pp. 378-394.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth (1996), "From data mining to knowledge discovery: An overview", *AI Magazine*, Vol. 17/3, pp. 37-54.

Greiff, S., and J. Funke (2009), "Measuring complex problem solving: The MicroDYN approach", in F. Scheuermann and J. Björnsson (eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing*, pp. 157-163, Office for Official Publications of the European Communities, Luxembourg.

Greiff, S., et al. (2013), "Complex problem solving in educational settings – Something beyond *g*: Concept, assessment, measurement invariance, and construct validity", *Journal of Educational Psychology*, Vol. 105/2, pp. 364-379, http://dx.doi.org/10.1037/a0031856.

Hall, M., et al. (2009), "The WEKA data mining software: An update", *ACM SIGKDD Explorations Newsletter*, Vol. 11/1, pp. 10-18.

Hämäläinen, W. and M. Vinni (2010), "Classifiers for EDM", in C. Romero, S. Ventura, M. Pechenizkiy and R.S.J.D. Baker (eds.), *Handbook on Educational Data Mining* , CRC Press, Boca Raton, FL, pp. 57-74.

Hershkovitz, A. and R. Nachmias (2011), "Online persistence in higher education web-supported courses", *The Internet and Higher Education*, Vol. 14/2, pp. 98-106.

Keim D.A. (2002), "Information visualization and visual data mining", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 8/1, pp. 1-8.

Mant, J. (2001), "Process versus outcome indicators in the assessment of quality of health care", *International Journal for Quality in Health Care*, Vol. 13/6, pp. 475-480.

Mazza, R. (2006), "Monitoring students in course management systems: From tracking data to insightful visualizations", *Studies in Communication Sciences*, Vol. 6/2, pp. 305-312.

Mazza, R. and V. Dimitrova (2003), "CourseVis: Externalising student information to facilitate instructors in distance learning", in U. Hoppe, F. Verdejo and J. Kay (eds.), *Proceedings of the International Conference in Artificial Intelligence in Education,* IOS Press, Sydney, pp. 279-286.

Mazza, R. and C. Milani (2005), "Exploring usage analysis in learning systems: Gaining insights from visualisations", paper presented at the Workshop on Usage Analysis in Learning Systems, 12th International Conference on Artificial Intelligence in Education, Amsterdam, 18 July 2005.

OECD (2014), *PISA 2012 Results: Creative Problem Solving (Volume V): Students' Skills in Tackling Real-Life Problems*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264208070-en.

OECD (2012), *Literacy, Numeracy and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264128859-en.

OECD (2010), *PISA 2012 Field Trial Problem Solving Framework*, draft subject to possible revision after the field trial, OECD, Paris.

OECD (2005), *Problem Solving for Tomorrow's World: First Measures of Cross-Curricular Competencies from* PISA 2003, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264006430-en.

Pechenizkiy, M., N. Trcka, E. Vasilyeva, W. van der Aalst, P. De Bra (2009), "Process mining online assessment data", in T. Barnes, M. Desmarais, C. Romero and S. Ventura (eds.), *Educational Data Mining 2009: 2nd International Conference on Educational Data Mining*, Cordoba, Spain, pp. 279-288.

Perera, D., J. Kay, I. Koprinska, K. Yacef and O. Zaiane (2009), "Clustering and sequential pattern mining of online collaborative learning data", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21/6, pp. 759-772.

Pfaff, Y. and F. Goldhammer (2012), "Evaluating online onformation: The influence of cognitive components", (manuscript submitted for publication).

Quinlan, J.R. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.

Romero, C., P.G. Espejo, A. Zafra, J.R. Romero and S. Ventura (2010), "Web usage mining for predicting final marks of students that use Moodle courses", *Computer Applications in Engineering Education*, Vol. 21/1, pp. 135-146, http://dx.doi.org/10.1002/cae.20456.

Romero, C. and S. Ventura (2010), "Educational data mining: A review of the state-of-the-art", *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, Vol. 40/6, pp. 601-618.

Romero, C., S. Ventura, P. Espejo and C. Hervas (2008), "Data mining algorithms to classify students", in *Educational Data Mining 2008, The 1st International Conference on Educational Data Mining*, Montreal, Canada, pp. 8-15.

Romero, C., S. Ventura and E. García (2008), "Data mining in course management systems: MOODLE case study and tutorial", *Computers & Education*, Vol. 51/1, pp. 368-384.

Talavera, L. and E. Gaudioso, E. (2004), "Mining student data to characterize similar behavior groups in unstructured collaboration spaces", paper presented at the Workshop on Artificial Intelligence in Computer Supported Collaborative Learning at the European Conference on Artificial Intelligence, Valencia, Spain.

Tóth, K., S. Greiff, S. Wüstenberg and H. Rölke (2014), "Discovering students' complex problem solving strategies in educational assessment", in: J. Stamper, Z. Pardos, M. Mavrikis and B.M. McLaren (eds.), *Proceedings of the 7th International Conferenceon Educational Data Mining*, pp. 225-228.

Tóth, K., H. Rölke and F. Goldhammer (2012a), "Investigating test-taking behaviour in simulation-based assessment: Visual data exploration", in *EDULEARN12 Proceedings CD*, 4th International Conference on Education and New Learning Technologies, Barcelona, 2-4 July 2012.

Tóth, K., H. et al. (2012b), "Investigating problem solving behaviour in simulated hypertext environments", unpublished manuscript.

Tóth, K., et al. (2012c), "Prediction of students' performance on test taking processes in Complex Problem Solving", paper presented at the 30th International Congress of Psychology, Cape Town, South Africa, 22-27 July 2011.

Tröka, N. and M. Pechenizkiy (2009), "From local patterns to global models: Towards domain driven educational process mining", in A. Abraham, J.M.B. Sánchez, F. Herrera, V. Loia, F. Marcelloni and S. Senatore (eds.), *Proceedings of the Ninth International Conference on Intelligent Systems Design and Applications,* pp. 1114-1119.

van der Aalst, W.M.P. (2011), "Process mining: Making knowledge discovery process centric", *ACM SIGKDD Explorations Newsletter*, Vol. 13/2, pp. 45-49.

van der Aalst, W.M.P. and A.J.M.M. Weijters (2004), "Process mining: A research agenda", *Computers in Industry*, Vol. 53/3, pp. 231-244.

van Dongen, B.F. et al. (2005), "The ProM framework: A new era in process mining tool support", in G. Ciardo and P. Darondeau (eds.), *Application and Theory of Petri Nets,* Springer, Heidelberg, pp. 444-454.

Wirth, J. and E. Klieme (2003), "Computer-based assessment of problem solving competence", *Assessment in Education: Principles, Policy, & Practice*, Vol. 10/3, pp. 329-345.

Zhang, H., et al. (2007), "Moodog: Tracking students' online learning activities", in C. Montgomerie and J. Seale (eds.*), Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications 2007,* AACE, Chesapeake, VA, pp. 4415-4422.

Zoanetti, N. (2010), "Interactive computer based assessment tasks: How problem-solving process data can inform instruction", *Australasian Journal of Educational Technology*, Vol. 26/5, pp. 585-606.

Chapter 13

EcoSphere: A new paradigm for problem solving in complex systems

By
David A. Tobinski
Department of Psychology, University of Essen, Germany.
and Annemarie Fritz
Department of Psychology, University of Essen, Germany.
Centre for Education Practice Research, University of Johannesburg, South Africa.

The western perspective on the environment has changed in the last decades. Many natural processes, such as hurricanes or tsunamis, and cultural ones, such as nuclear power plants or space missions, hold immense potential to cause problems, so the current state of the world is often seen as a complex system. Dealing with complex systems, particularly unstable ones with autonomously changing (eigendynamic) elements, will be one of the key competences for future generations, which is why measuring and training these competences is so important today. This chapter focuses on measuring competency at problem solving in complex systems (PSCS). It introduces the processes involved in human problem solving – from observation, whether through instruction or active construction, to exploration and finally regulation, and discusses their relations to the environment, which is built up of complex systems. It then presents the new assessment tool EcoSphere, which is a simulation framework for testing and training human behaviour in complex systems. In particular, it allows test takers' previous knowledge to be taken into account to better measure their problem-solving abilities.

The NEPS data collection is part of the Framework Programme for the Promotion of Empirical Educational Research, funded by the German Federal Ministry of Education and Research, and supported by the Federal States.

Introduction

The development and evolution of mankind's perspective on our environment go hand in hand with the handling and forming of this environment. It is a cycle whose elements are difficult to differentiate: a changing environment will lead to new behaviour and this leads to new perspectives on the environment, while new perspectives lead to new behaviour, which will change the environment. It is clear that complex systems have evolved from technical progress such as constructing the space shuttle. On the other hand, complex systems in nature, like the global climate system, have existed for billions of years but can only now be understood, since human perspectives have developed together with the scientific knowledge in the last centuries (Weaver and Rusk, 1958; Whitehead, 2001). The human perspective on our environment here stands for the entire human cognitive architecture embedded in our environment. One of the most interesting questions in this circle is: what changes the human perspective of its environment? In our view there is only one answer: a solved problem.

The EcoSphere

The new assessment tool *EcoSphere* has some well known and very interesting roots. So what does it stand for? The prefix "eco" is most commonly encountered as part of the words ecology and economy. In Western society, these words were first popularised by the book *Walden or Life in the Woods* (1854) by Henry David Thoreau (1817-62). "Eco" originally comes from the Greek *oikos*, which means house. This metaphor of the house stands for all objects or variables and their relations within the house. This, incidentally, is also a good metaphor for the systems involved in problem solving in complex systems (PSCS) and dynamic decision making (DDM). The word "sphere" comes from the Greek *sphaira*, which means ball or globe. We will see this later in the first scenario of the EcoSphere, which is called the BioSphere.

Why a new assessment tool?

Why do we need a new assessment tool for problem solving? The origins of complex problem-solving research started with huge and impressive systems like Lohhausen and Moroland (Dörner, et al., 1983; Brehmer and Dörner, 1993). The demand for more systematic designs (Hussy, 1985) brought a second phase with smaller systems like DYNAMIS (Funke, 1993). Today we find a new and very clear approach using micro-systems like MicroDYN (Wüstenberg, Greiff and Funke, 2012). This trend of reducing complexity has led from research into complex problem solving (CPS) to research into dynamic problem solving (DPS). Many difficulties with measuring problem-solving abilities have been recognised and the considerable problem of one-item testing seems to be elegantly mastered by using/creating smaller items (Greiff and Funke, 2009). However, research into CPS started with the question of ecological validity (Brehmer and Dörner, 1993) and there are still open and even unseen questions on the horizon: our big problems in nature and society mostly arise from oscillating systems with eigendynamics (see below), like ecological and economic systems. Yet where are the systems with eigendynamics in CPS research? While there are real-time simulations available for modelling, such as NetLogo (Wilensky, 1999), no system with real-time simulation is available for assessment, experimental or intervention studies. Second, a problem is distinguished from a task by missing knowledge, which depends on the individual knowledge of the person at the start of the problem-solving process. This missing knowledge does not just vary dichotomously. This central theoretical fact has not been considered empirically until now: is it possible to to measure previous knowledge in CPS? The following sections will give offer some reflections on these important questions.

Eigendynamics

In EcoSphere the concept of eigendynamics (EDY) follows the understanding of natural dynamics (Christophersen, 1999; Bossel, 2004; Gobert et al., 2011). A lot of mathematical work has been done on this topic in the last decades (Kauffman and Varela, 1980). At the core of eigendynamics is the concept of values changing without any interaction: the system is changing itself continuously. While eigendynamics depends on the way the variables are connected to each other and themselves, the most important factor is time. Systems with real eigendynamics show a fluid behaviour over time. While most dynamic systems in use may show increasing or decreasing values, systems with EDY are able to show oscillating values. This is the most interesting fact about EDY for problem-solving research, the characteristic of being observable without interaction, i.e., being produced automatically. The next section takes a deeper look at the process of observation and the other central information processes in the EcoSphere.

Previous knowledge

The product of successful problem-solving processes can be seen in the construction of knowledge.This constructive process starts from a point between knowledge and missing knowledge: a problem never appears in the middle of knowledge-free spaces but at their borders. Thus, the initial state of a problem is based on previous knowledge (K_p). The distinction between a task and a problem derives from previous knowledge, which can be a spectrum and not just a binary dimension. In understanding problem solving as information processing, the relation between a system and previous knowledge gives us two new indexes. First, previous knowledge gives us the ability to measure the individual problem complexity at the start of the problem. Using a cybernetic concept of information (von Cube, 1968; Schweitzer, 1999) we measure the individual problem complexity as information weight (I_w) which is the difference between the structural information of the system (${}_sI_{stc}$) and the information contained in the previous knowledge (${}_pI_{Kp}$) of the person: $I_w = {}_sI_{stc} - {}_pI_{Kp}$. The second measure is the index of pragmatic information, which is the difference between the previous knowledge at the initial state and the structural knowledge at the goal state, or the successfully processed information used to solve the problem. Thus, previous knowledge becomes a central part of analysing problem-solving processes; it is needed to simulate ecologically valid semantic scenarios. In spite of using artificial semantics, EcoSphere uses real-life and scientific curriculum-based semantics, to give participants the opportunity to apply their previous knowledge more precisely. When measuring problem-solving abilities, strategies can thus be better differentiated and new types of strategies can be identified.

Information processing in EcoSphere

On regulating the environment, Conant and Ashby wrote, “Every good regulator of a system must be a model of that system” (1970: 89). Looking more closely at that sentence, we realise that the authors are describing the point *after* complex problem solving. The end point after successful problem solving is the ability to be good regulator. Before then, people might be good or bad regulators or even non-regulators, but they are all in the position of an observer and there will be no non-observers.

Observation

The starting point of problem solving is always connected with observation (Maturana, 2000). Imagine, for example, that your car is making unusual sounds. You have the ability to interpret the sound as unusual, which your passenger might not even realise. So he has no problem until that point, as he is observing your car from the basis of his own abilities. Observing means that an observer is able to identify entities in the current information flow, which is only possible by differentiating one entity from another and realising their relationships (Luhmann, 2001; Maturana, 2000; Glasersfeld, 1996). To your passenger, the sound of your car is a variable with a wide variance, but you can differentiate two variables: the variance of good motor sound and the variance of bad motor sound, which might be followed by engine damage. This is an example of interpretation, which is seen as a process of comparing current information with previous knowledge. The previous knowledge has to be reconstructed and the outcome of that comparison will update the current mental model of the observer (Buckley et al., 2010).

While your passenger is observing the sound of your car, he is actively constructing an instance from the selected and given information and his own observing abilities (Luhmann, 2001; Gonzalez et al., 2003). Differentiating the various motor sounds of your car lies beyond his observing abilities. All construction processes are similar to the processes of instruction until that point (Cobern et al., 2010). Both instruction and construction depend on the observing abilities of an information receiver, respectively an observer. The difference between them lies in the active selection of information from the environment in the case of construction, and the passive receiving of information in the case of instruction. A constructive situation presents all the information in parallel to the receiver, while an instructive situation always presents the information in a serial way. In the constructive situation, the receiver might miss a lot of information because their observation strategies are insufficient. In contrast, the instructed person only has to pay attention to the guiding media.

It is important to note that systems are only observable if they have the attribute of eigendynamics, otherwise they would only be describable in their structure and not in their processes. EcoSphere offers both instructional and constructional situations for experiments in the same semantic content. Being able to measure the observing abilities of a participant is an advantage for measuring their problem-solving abilities, because observing is part of that construct. However, the EcoSphere program goes beyond those aspects, because the "magic moment" of problem solving lies beyond these discussed points. Our central question is: what changes the observing abilities of an observer?

Exploration

We can see from our example above that the quality of observation depends on the ability to interpret different variables in a given spectrum of variance. But how do we identify variables and their attributes? What causes one entity to be split into two or more entities? Consider a second example: thousands of years ago, our ancestors loved to eat berries. The berries' diverse colours and size alone was no encouragement to differentiate them on a deeper level or to give them special names. Humans tried to eat every berry until someone tried a "special" one. Before then, people might never have expected a lethal berry or in fact any lethal food at all, until that first unexpected error occurred. To encode this deadly error beyond a single individual's mind, the culture might give those instances of lethal berries a special name, such as "hell berries". The observing ability regarding berries deepened and the single entity of berry was split into edible and lethal ones. Culture was now able to instruct its members to avoid lethal berries.

Central to this example is the need to learn from error to become a better observer. This recalls the definition of problem solving as "[…] varying mixtures of trial and error and selectivity" (Simon, 1962:472). The first step is to realise the possibility of an error, which leads to hypotheses about the space, and the second step is to risk an error. In our example, humanity was not able to think about the possibility that there might be any lethal food. After that lethal experience the new knowledge was also transferred to other foods. People started to explore the environment for edible food and formulated a lot of hypotheses. Exploration represents the act of trial-and-error: behaviour producing hypotheses, testing hypotheses and checking the results within a problem space, an evaluation, which always allows the production of errors. Experience is central to the construction of new observing abilities and growing opaqueness of the system, which is deeply connected to insight and awareness (Sternberg and Davidson, 1995). Thus, the code of language is helpful for transmitting important and lifesaving information, but the first information about lethal food cannot be constructed by language alone, it has to be generated through experience and then bound to language.

In our century of computer-based learning, we are able to simulate explorable environments with dangerous aspects while protecting learners from harm. The simulation of a system allows its deconstruction without any risk. Explorable environments provide freedom to vary parameters using the exogenous system variables. EcoSphere simulates complete environments for this central phase of problem solving. However, complex problem solving does not reach its culmination until goal-driven behaviour begins. This is the point at which the problem solver slowly leaves the problem space and is becomes a regulator in a planning space.

Regulation

Once again we return to the words of Conant and Ashby, (1970: 89): a model of the system implies a deep knowledge about the system. To finally close the cycle of problem-solving processes, we look again at our first example. The first experience of the unusual sound of the car was followed by engine damage. Years later, after the engine had been repaired, the sound occurred once more. While our passenger was not able to differentiate the sounds of the car, we remembered and the engine broke down again. Yet our power of observation increased: we realised that the first problem occurred after 10 000 miles and this second breakdown after 20 000 miles. In addition, our mechanic told us that both times the car had run out of oil. From that day, we were able to regulate the engine of our car and refill its oil every 9 999 miles and the problem never recurred. Our example shows that regulation can be seen as the act of keeping a dynamic system in a stable state. This state can be fixed or vary within a specific range, but we need to keep it stable by applying instructed or constructed knowledge. We can also see that the action is bound to prognostic abilities about the systems states: the regulator knows when best to regulate the system. Sometimes it is important to combine a lot of different operations to get a system on the right track, but it is always connected to processes of anticipation and prognosis. Otherwise, it would be successful trial-and-error behaviour. When the person has learned to be a model of the system, we say that the problem space has been transformed into a planning space and the problem solver has become a planner. Table 13.1 summarises the different phases of human-system interaction and the processes involved.

Table 13.1. **Information processing and interaction with the system**

	Instructed system	**Constructed system**	**Exploring the system**	**Regulating the system**
Information processing	receive	select	select	select
	interpret	interpret	interpret	interpret
	encode	encode	encode	encode
	re-construct	re-construct	re-construct	re-construct
			deconstruct	deconstruct
			evaluate	evaluate
				apply
Human-system interaction	non interaction	non interaction	parameter-changing	parameter- changing
				goal-driven parameter-adjusting

The EcoSphere program: BioSphere scenario

The EcoSphere software consists of different modules: the introductory module, the instruction module and the exploration module. The example of the first scenario, BioSphere, shows these modules at work. BioSphere is influenced bysystems which were developed by the National Aeronautics and Space Administration (NASA) in the research programme "Controlled Ecological Life Support System" (CELSS) in the early 1980s (Cullingford, 1981; Macelroy, 1990).

Introductory module

Modern computer games motivate their users by placing them in a story. EcoSphere adopts a similar approach to keep participants on track. For this, a storyboard has been developed and cartoon characters have been created. While the software as a whole is introduced by a general introduction, each individual scenario gets its own scenario introduction.

Storyboard

EcoSphere is set in a laboratory in the near future. The participant is a member of a research team who have to solve different scientific problems. The team consists of a professor, a female and a male assistant, and a dog and a robot. Considering participants' motivation and interests, many different graphical alternatives for these characters and their surroundings have been developed. Studies of 55 9th and 10th grade students identified the most liked subjects and objects , which have been integrated into the software.

General introduction

The general introduction presents the scientific team, their laboratory and the surroundings. It is also a tutorial for the general handling of the software and its functions. The most important function is that of drawing computer-based causal diagrams, which will be discussed below. Taking the graphical appearance of modern computer games into account, the general introduction was animated using the professional high-end computer-animation software ToonBoom, which is also used for professional productions like the TV show *The Simpsons*. This part of the software is enabled to use speech.

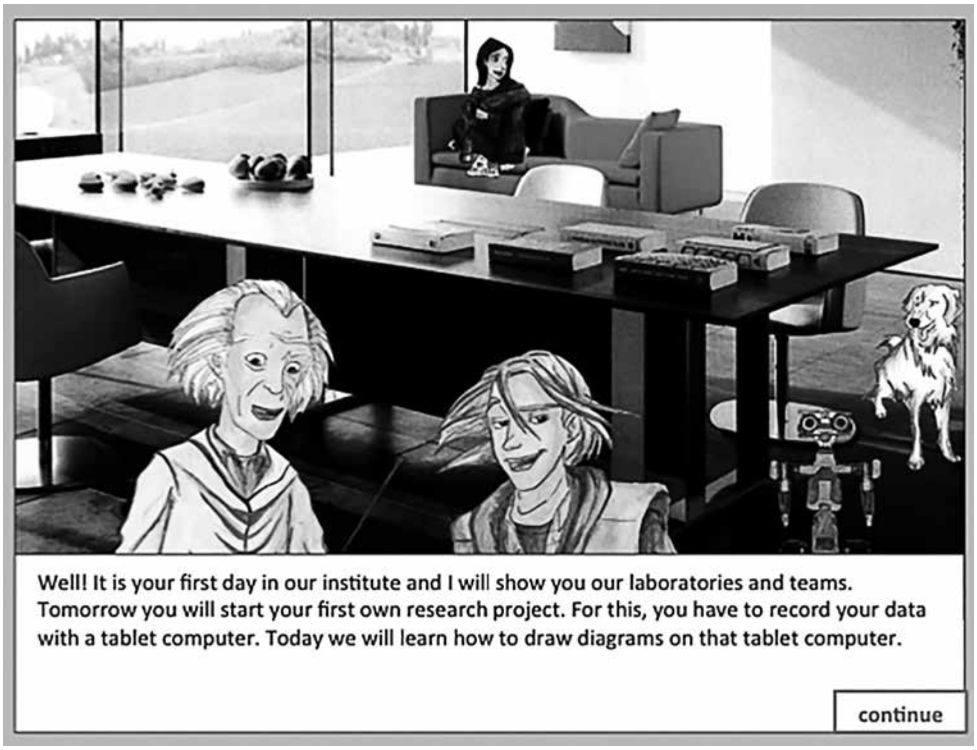
Scenario introduction

The scenario introduction introduces the plot of a scenario and presents the given state of the problem with all the important information. In contrast to the general introduction, the scenario introduction is editable. Visual instructions can be created as well as audio instructions about modality effects. In the BioSphere scenario, the scientific team has received a parcel from NASA with a small ecosystem in it. The briefing can now be adjusted for different experimental designs: 1) observing the ecosystem; 2) exploring the ecosystem; or 3) regulating the ecosystem. A new approach in this field is to include instructed systems knowledge. In this, participants are given complete information about the ecosystem without any problem-solving processes: they will obtain the information of a full instructed ecosystem (see Figure 13.1).

Instruction and construction module

The second module (ECO-I-C) contains the phases of instruction and construction. Both phases are automatically followed by a test of the declarative knowledge, which is measured via causal diagrams. EcoSphere uses a new standard for drawing computer-based causal diagrams (CBCD).

Figure 13.1. **Screenshot of a scenario introduction**
The characters introduce participants to the scientific team



Well! It is your first day in our institute and I will show you our laboratories and teams. Tomorrow you will start your first own research project. For this, you have to record your data with a tablet computer. Today we will learn how to draw diagrams on that tablet computer.

Instruction

The instruction phases can be used for two different purposes. First, it can give information to participants similar to an operating manual. This is the way it is used in the introduction module. Second, it can provide the complete information of any known describable system. instructions are mostly given in writing, but the EcoSphere software also enables audio instructions. If the design of a study uses this second function, participants receive complete information and only have to interpret and encode it. In the BioSphere scenario, 40 local, stochastically independent items have been developed for instructed complex systems. These items range from small systems with two variables and one connection, to bigger systems with five variables and five connections. The content is taken from different domains to consider possible semantic effects. A first study in our lab, still unpublished (*N = 200*), could highlight the effects between the number of variables and the number of connections and show the possibility of comparing information processing in instructed, constructed and explored systems.

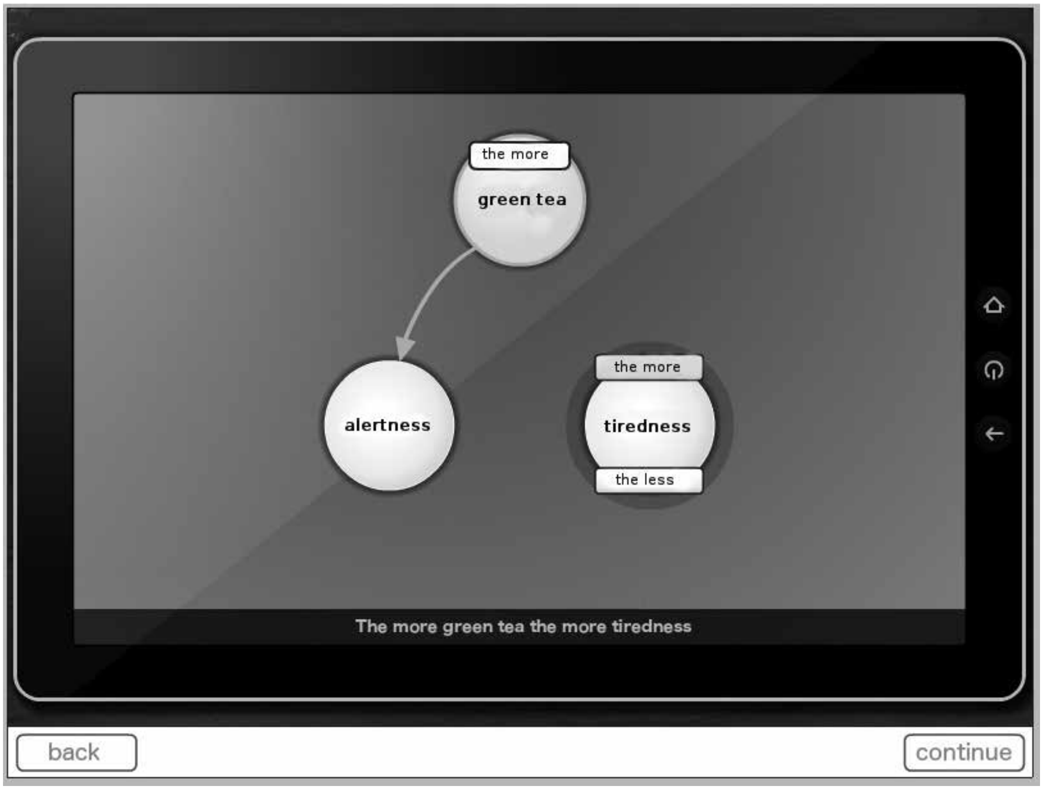
Construction

The construction phase provides information about the system in a visual way. The core of the EcoSphere software is to simulate observable and explorable systems. While the participant can interact with the system in the exploration phase, discussed in the next section, in the construction phase, they are presented with the behaviour of an observable system as an animation, which in contrast is non-interactive. Observable systems can be implemented as ActionScript files (swf), other files such as animated gifs, or movie files (mp4). The BioSphere scenario presents an ecosystem with an animal population and a plant population to the participant, as described below.

Computer-based causal diagrams

The instruction and construction phases have no interaction between the person and the system, but, based on participants' observation abilities, they will modify the person's declarative knowledge. Particpants encode an instance of the particular system, which can be reconstructed and drawn in a computer-based causal diagram (CBCD). In addition to context-action exemplars, we also regard context-observing exemplars as discrete representations of knowledge and see them as "instances" (Gonzalez, 2012). Drawing a CBCD after an instruction or construction phase involves multiple decisions after description. Before the participants draw a CBCD of an instructed or explored system, they first go through training on CBCD for three items (see Figure 13.2).

Figure 13.2. **The interface for drawing computer-based causal diagrams**



The more green tea the more tiredness

CBCDs represent the system in a specific syntax. At the start, the variables are presented to the participant, who have to draw the direction and the positive or negative values of the connections. While the participant is composing the connections, the software is also translating the syntax of CBCD into sentences. For example, if the topic of an item was a glacier system a connection could be stated as: "The more sunshine, the less ice on the glacier". Taking the flooding of working memory into account, participants have the option to look at a notepad, an external memory which contains the instruction. Any such fallback to instructional information can be measured, including how often and how long the information was accessed. The most interesting benefit of EcoSphere is the ability to take the previous knowledge of the participants into account and obtain measures for information weight and pragmatic information. This previous knowledge can also be extracted from the CBCDs. Taking previous knowledge into account is a new way of research in problem-solving measurement, which will be examined in the next section.

Exploration and regulation module

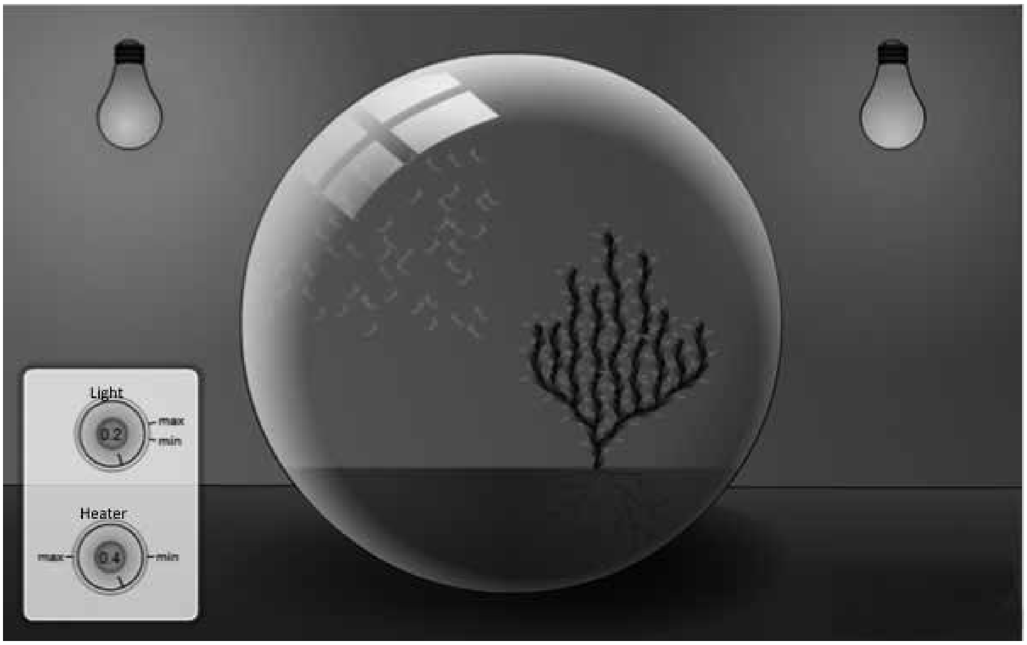
The core of problem solving in complex systems is the processes of exploration and regulation. EcoSphere uses a third module (ECO-E-R) to support the diagnosis of exploration and regulation abilities.

Simulation

The ECO-E-R module of EcoSphere is where the interactive simulations take place. The aspect of ecological validity is important for simulations, which was the claim in early research into complex problem solving and sometimes seems to be forgotten in current approaches (Brehmer and Dörner, 1993). The deeper sense of using ecologically valid scenarios can be seen in the new approach of regarding problem solving as the construction of new knowledge within previous knowledge. Previous knowledge is regarded as a prerequisite in PSCS. Participants have to draw a CBCD before they go into the exploration phase.

The simulation itself presents an environment with different exogenous and endogenous variables, also called stocks in system dynamics (Forrester, 1968; Bossel, 2004). All environments are predesigned using the system dynamics modeller of the NetLogo software (NetLogo 5.0.4; Wilensky, 1999). Simulations in EcoSphere have to produce eigendynamics, which leads to continuously changing behaviour. Only real eigendynamics allow observation without participant interaction. Oscillating systems are very useful for this approach. In EcoSphere, oscillating systems can be presented in stable or instable conditions. After successfully designing those simulations and their underlying differential equations, all will be programmed in ActionScript (2.0 and 3.0) for using a flash format.

Figure 13.3. **The first BioSphere scenario consists of two exogenous and three endogenous variables**



Observableness

The BioSphere scenario simulates a small ecosystem with an animal and a plant population. The exogenous variables of light and temperature can be varied in their intensity by control dials, allowing participants to explore the system and observe the reactions. The editor of the secnario can pre-assign the grade of "observableness", which is connected to the level of knowledge that can be extracted by the user (Wilensky and Resnick, 1999). For example, the system can respond to changes immediately or with a predefined delay; the reaction can be given in different codes, graphical and/or numerical (Gobert et al., 2010). If the graphical observableness in the BioSphere scenario is set to detailed, the animal population responds to changing temperatures with changing coloration: the yellow animals change to red or blue. When temperatures are too warm or too cold, the animals will not survive. In other settings, only the dying would be observable, so the correlation between the temperature and the colour of the animals would not be as obvious. In this way, even distant effects and side effects within the system can vary in their evidence. While those settings are defined as intrinsic memory load (Sweller, 1994), EcoSphere is able to use some effects that might motivate participants, but can be controlled as extraneous load. For example, the swimming behaviour of the animal shoal can be made to look very vivid and natural or more mechanical and clear.

In EcoSphere, errors, (respectively deconstruction) lead to a decrease of the animal or plant population or both, or even to their extinction. While the exploration phase of EcoSphere is seen as a chance to learn from errors and to construct a model of the system, during the regulation phase these errors should ideally not occur or at least not end in the death of a population.

Regulation

While the exploration phase offers problem-based learning with the aim of constructing a complete mental model of the system, the regulation phase challenges the participant to apply this new system knowledge to reach a given goal. The regulation phase therefore always starts with a system in an unstable state and the participant has to get the system into a stable state. In the current systems, there is always only one perfect combination of parameters that will lead to the stable system. The number of attempts the participant is allowed to make can be predefined in the regulation phase. For example, the system might close after 20 trials and give feedback about how many days the system would have survived with the last parameters for light and temperature. Or participants could end the regulation phase before the 20th trial if they supposed the current setting was the perfect one. If so, they would have produced an infinitely stable system, a result that is not ecologically valid from the current perspective of physics. Maybe a computer simulation is more metaphysical than we expect.

Outlook and conclusion

Outlook

The technical aspects of EcoSphere are moving forward. The software is designed to operate both offline and client-based as well as server-based. For larger assessments, log files are managed on secure servers and the log-file analysis will be automated using R on the server side to give faster feedback to the institutions. The client-based version of EcoSphere has been tested successfully on tablet computers. Within the modules ECO-I-C is going to be prepared for adaptive testing.

Conclusion

Current studies in complex problem-solving research focus on the correctness of the solvers' mental model and control performance as the two main indicators. Both are enhanced in the EcoSphere by taking into account the previous knowledge of the participant. In addition, EcoSphere offers more precise analysis of the control phase to interpret more different problem-solving strategies. For this type of "what-if" knowledge, the authors use the term imperative knowledge

(Abelson et al., 1996) in contrast to non-declarative procedural knowledge. Analysing and contrasting the four types of interaction with systems – instruction, construction, exploration and regulation – promises an exciting way to observe a person transforming from a problem solver in complex systems to a planner in complicated systems.

References

Abelson, H., G.J. Sussman and J. Sussman (1996), *Structure and Interpretation of Computer Programs*, MIT Press, Cambridge, MA.

Bossel, H. (2004), *Systeme Dynamik Simulation*, Books on Demand, Norderstedt.

Brehmer, B. (1995), "Feedback delays in complex dynamic decision tasks", in P.A. Frensch and J. Funke (eds.), *Complex Problem Solving: The European Perspective*, Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 103-130.

Brehmer, B. (1992), "Dynamic decision making: Human control of complex systems", *Acta Psychologica*, Vol. 81/3, pp. 211-241.

Brehmer, B. and D. Dörner (1993), "Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study", *Computers in Human Behavior*, Vol. 9/2-3, pp. 171-184, http://dx.doi.org/10.1016/0747-5632(93)90005-D.

Buckley, B., J. Gobert, P. Horwitz and L.M. O'Dwyer (2010), "Looking inside the black box: assessing model-based learning and inquiry in BioLogica™", *International Journal of Learning Technology*, Vol. 5/2, pp. 166-190.

Christophersen, R.G. (1999), *Geosystems: An Introduction to Physical Geography*, Pearson Education, London.

Cobern, W.W. et al. (2010), "Experimental comparison of inquiry and direct instruction in science", *Research in Science & Technological Education*, Vol. 28/1, pp. 81-96.

Conant, R.C. and W.R. Ashby (1970), "Every good regulator of a system must be a model of that system", *International Journal of Systems Science*, Vol.1/2, pp. 89-97.

Cullingford, H. (1989), "Development of the CELSS Emulator at NASA JSC", in *Controlled Ecological Life Support Systems*, NASA, Washington, Columbia, pp. 319-326.

Dörner, D., H.W. Kreuzig, F. Reither and T.H. Stäudel, T.H. (1983), *Lohhausen: Vom Umgang mit Unbestimmtheit und Komplexität*. Huber, Bern.

Dörner, D. (1979), *Problemlösen als Informationsverarbeitung* [Problem Solving as Information Processing], Kohlhammer, Stuttgart.

Forrester, J.W. (1968), *Principles of Systems*, 2nd edition, Pegasus Communications, Waltham, MA.

Frensch, P.A. and J. Funke (1995), "Definitions, traditions, and a general framework for understanding complex problem solving", in P.A. Frensch and J. Funke (eds.*), Complex Problem Solving: The European Perspective*, Erlbaum, Hillsdale, NJ, pp. 3-25.

Funke, J. (2003). *Problemlösendes Denken* [Problem Solving and Thinking], Kohlhammer, Stuttgart.

Funke, J. (1993), "Microworlds based on linear equation systems: A new approach to complex problem solving and experimental results", in G. Strube and K.-F. Wender (eds.), *The Cognitive Psychology of Knowledge*, Elsevier Science Publishers, Amsterdam, pp. 313-330, http://dx.doi.org/10.1016/S0166-4115(08)62663-1.

Funke, J. (1990), "Systemmerkmale als Determinaten des Umgangs mit dynamischen Systemen" [System attributes as determinants of performance in dealing with dynamic systems], *Sprache & Kognition*, Vol. 9/3, pp. 143-154.

Funke, J. and P.A. Frensch (2007), "Complex problem solving: The European perspective – 10 years after", in D.H. Jonassen (ed.), *Learning to Solve Complex Scientific Problems*, Erlbaum, New York, pp. 25-47.

Glasersfeld, E. von (1996), *Radikaler Konstruktivismus* [Radical Constructivism], Suhrkamp, Frankfurt am Main.

Gobert, J., A.R. Pallant and J. Daniels (2010), "Unpacking inquiry skills from content knowledge in geoscience: A research and development study with implications for assessment design", *International Journal of Learning*, Vol. 5/3, pp. 310-334.

Gobert, J. et al. (2011), "Examining the relationship between students' understanding of the nature of models and conceptual learning in biology, physics, and chemistry", *International Journal of Science Education*, Vol. 33/5, pp. 653-684.

Greiff, S., and J. Funke (2009), "Measuring complex problem solving: The MicroDYN approach", in F. Scheuermann and J. Björnsson (eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing*, pp. 157-163, Office for Official Publications of the European Communities, Luxembourg.

Gonzalez, C. (2012), "Training decisions from experience with decision making games" in P.J. Durlach and A.M. Lesgold (eds.), *Adaptive Technologies for Training and Education*, Cambridge University Press, Cambridge, pp. 167-178.

Gonzalez, C., J.F. Lerch and C. Lebiere (2003), "Instance-based learning in dynamic decision making", *Cognitive Science*, Vol. 27/4, pp. 591-635, http://dx.doi.org/10.1016/S0364-0213(03)00031-4.

Gonzalez, C., P. Vanyukov and M.K. Martin (2005), "The use of microworlds to study dynamic decision making", *Computers in Human Behavior*, Vol. 21/2, pp. 273-286, http://dx.doi.org/10.1016/j.chb.2004.02.014.

Hussy, W. (1998), *Denken und Problemlösen* [Thinking and Problem Solving], Kohlhammer, Stuttgart.

Hussy, W. (1985), "Komplexes Problemlösen – Eine Sackgasse?" [Complex problem solving: A dead end?], *Zeitschrift für Experimentelle und Angewandte Psychologie*, Vol. 32, pp. 55-74.

Hussy, W. (1984), *Denkpsychologie* [Psychology of Thinking], Kohlhammer, Stuttgart.

Kauffman, L.H. and F.J. Varela (1980), "Form dynamics", *Journal of Social Biological Structures*, Vol.3/2, pp. 171-206.

Luhmann, N. (2001), *Aufsätze und Reden* [Papers and talks], Reclam, Stuttgart.

Macelroy, R. (1990), "Controlled ecological life support systems", CELSS'89 Workshop, NASA, Washington, Columbia.

Maturana, H.R. (2000), *Biologie der Realität* [Biology of reality], Suhrkamp, Frankfurt am Main.

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

Newell, A. and H.A. Simon (1961), "Computer simulation of human thinking", Science, Vol. 134/3495, pp. 2011-2017.

Schweitzer, F. (1999), “Structural and functional information: An evolutionary approach to pragmatic information”, in W. Hofkirchner (ed.), *The Quest for a Unified Theory of Information*, Gordon and Breach, Amsterdam.

Simon, H.A. (1962), “The architecture of complexity”, *Proceedings of the American Philosophical Society*, Vol. 106/6, pp. 467-482.

Sternberg, R.J. and J.E. Davidson (eds.) (1995), *The Nature of Insight*, MIT Press, Cambridge, MA.

Sweller, J. (1994), “Cognitive load theory, learning difficulty and instructional design”, *Learning and Instruction*, Vol. 4/4, pp. 295-312.

Sweller, J., and S. Sweller (2006), “Natural information processing systems”, *Evolutionary Psychology*, Vol. 4, pp. 434-458.

Thoreau, H.D. (1995), *Walden or Life in the Woods*, Dover Publications, New York.

von Cube, F. (1968), *Kybernetische Grundlagen des Lernens und Lehrens* [Cybernetic sources of teaching and learning], Klett, Stuttgart.

Weaver, W. and D. Rusk (1958), *A Quarter Century in the Natural Sciences*, New York.

Whitehead, A.N. (2001), *Denkweisen* [Thinking styles], Suhrkamp, Frankfurt am Main.

Wilensky, U. (1999), “NetLogo”, Center for Connected Learning and Computer-Based Modeling, Northwestern University http://ccl.northwestern.edu/netlogo/.

Wilensky, U. and M. Resnick (1999), “Thinking in levels: A dynamic systems approach to making sense of the world”, *Journal of Science Education and Technology*, Vol. 8/1, pp. 3-19.

Wüstenberg, S., S. Greiff and J. Funke (2012), “Complex problem solving: More than reasoning?”, *Intelligence*, Vol. 40/1, pp. 1-14, http://dx.doi.org/10.1016/j.intell.2011.11.003.

PART V

Future issues: Collaborative problem solving

Chapter 14

Assessment of collaborative problem-solving processes

By
Esther Care
Brookings Institution, Washington, United States.
and Patrick Griffin
Assessment Research Centre, University of Melbourne, Australia.

This chapter presents a framework for understanding collaborative problem solving, including both the cognitive and social perspectives, and identifies the construct's theoretical underpinnings, structure and elements. It describes the circumstances under which collaborative problem solving might best be used, with consequences for the design of tasks to assess the component skills. It highlights the characteristics of problem-solving tasks, interdependence between problem solvers and the asymmetry of stimulus and response, through a focus on task design. The chapter then outlines approaches to measuring collaborative problem solving and illustrate them with examples.

The authors wish to acknowledge the contributions of the ATC21S expert panel on collaborative problem solving, led by Friedrich Hesse and Eckhard Klieme, and including Esther Care, Marlene Scardamalia and Kurt Van Lehn.

Introduction

The identification of 21st-century skills is a topic which has attracted a great deal of global attention since the turn of the century. This is shown by the establishment of many national and international consortia (such as the Partnership for 21st Century Skills), research endeavours such as the MIT Center for Collective Intelligence, teacher forums such as the Teacher Leaders' Network, and government initiatives such as the National Research Council of the National Academies). There have been many different approaches to the topic with diverse areas of focus such as labour needs, e-learning spaces, digital technologies and their application in the traditional classroom, and the solutions to important and global problems.

The approach to defining, describing, measuring and assessing collaborative problem solving outlined in this chapter derives from the Assessment and Teaching of 21st Century Skills (ATC21S) project, 2009-12. The project's goal was to scope the framework of 21st-century skills, and to develop prototype technology-based assessment and teaching tasks which, combined with appropriate reporting methods and professional development guides for teachers, could be used to inform teaching in the classroom. The focus was on new ideas, new approaches to assessment, new skills and assessing existing skills in different ways. The project was a partnership between the private sector (Cisco, Intel, and Microsoft), the public sector (the governments of Australia, Costa Rica, Finland, the Netherlands, Singapore and the United States), and academia (University of Melbourne).

Those who are concerned with how teaching and learning may look in the future often identify strategies such as increasing openness to new ideas, approaching the future with confidence, being accepting of change and having "vision", but these affective approaches ignore the need to identify and enhance relevant skills. Increasing dependence on information technologies and changing social trends in the use of technology to create and use information bring into focus the need to teach students how to manipulate these technologies and how to navigate the communication channels available to them.

Organisational structures, and the roles played by workers within them, have changed. This has led to changing work techniques, with many being affected by newly available information and communications technology (ICT) infrastructure. Work in such changed environments requires different skill sets. There are varied approaches to investigating the skill sets required by current work environments. One approach looks at the workforce implications, where automation has led to a diminished need for humans to perform repetitive and routine tasks. In such workplaces, workers need a much more sophisticated skill set including complex problem solving and co-ordination skills. However, it is not clear whether such skill sets are new, simply need to be applied to a new work environment, or are merely needed much more than before.

Countries have shifted more of their labour capacity to knowledge and information production and transfer, creating the need for education systems to equip their student populations with a new generation of skills beyond literacy and numeracy. Autor, Levy and Murnane (2003) provide a compelling example of how in the last three decades of the 20th century demand has moved from routine cognitive tasks such as bookkeeping and filing, and from routine manual tasks such as assembly line work, to tasks that require expert thinking, such as identifying and solving new problems, and tasks that require complex communication. The use of the term "expert" is clarified by Murnane (National Research Council, 2011) as referring to:

- deep understanding of the particular domain and relationships within it
- pattern recognition
- a sense of initiative
- metacognition (i.e., monitoring your own problem solving)
- complex communication and its components.

This set of descriptors brings Murnane's thinking on expertise very close to the construct of collaborative problem solving, generalising this specific skill construct to a set of skills essential for functioning in the 21st century.

It has been argued (Anderman, 2011; e.g. Kuncel, 2011, both published in National Research Council, 2011) that skills such as critical thinking are not domain-general but domain-specific, and that there are problems with the transfer of skills from one context to another. This issue is relevant if we are to substantiate the promotion of collaborative problem solving as a teachable and learnable set of skills. There is evidence that students may become more adept in a specific skill if it is taught, but this does not necessarily generalise across contexts. So when general problem-solving strategies are taught, they should be taught within meaningful contexts and not simply as rote algorithms to be memorised. Students need to be taught explicitly how to transfer skills from one context to another so that they can recognise that the solution to one type of problem may be useful in solving a problem with similar structural features (Mayer and Wittrock, 1996).

The description of collaborative problem solving outlined in this chapter concentrates on the application of 21st-century skills, with their dependence on electronic media for communication. This context imposes certain characteristics on collaborative problem-solving activities, which need to be reflected in how they are taught and learned. The assessment approach outlined here epitomises this perspective. It specifies that collaborative problem solving has a number of components: collaboration, problem solving and an understanding of relevant new technology. The skills should be useful in everyday life, as well as in education and employment.

Collaborative problem-solving skills: A framework for understanding

The amount and diversity of knowledge or information needed to solve a problem, and whether the problem has a finite solution or not, may determine whether it can be solved by a single individual. Where one individual has the capacity to solve a problem, and where this is an efficient process, there is little point in working collaboratively – if the focus is on reaching a solution. If the focus is on other goals, such as facilitating group processes or teaching individuals – students or workers – to work together, collaborative problem-solving activities may provide a useful teaching or training tool. Problem-solving strategies vary depending on the problem. For example, "means-end analysis" (Newell, 1963) is a common strategy for solving simple, formalised problems. More complex problems require different analytic approaches that organise multiple inputs, their combination, and the rules governing this process. The demands of solving a particular problem depend on how much and what type of knowledge is needed, the complexity of the problem space, and the skills of those trying to solve the problem.

Collaborative problem solving, one of the 21st-century skills (identified in Binkley et al. (2010), is a complex process that requires participants to externalise their individual problem-solving processes and to co-ordinate these contributions into a coherent sequence of events (Hesse et al., 2015). In many respects, it is not the same as individual problem solving (Campbell, 1968), not least because of its requirement for communication and interaction. This factor goes to the heart of the rationale for identifying collaborative problem solving as a skill in its own right – that contributions from more than one person are required. Participants are normally assumed to be human persons (see below for a discussion of the impact of introducing digital agents) and, in this discussion, assumed to be students who are learning to interact collaboratively to solve problems. We assume that collaboration demands the interaction of two or more students. The context has implications for the range of subskills that are required. Students may interact face-to-face, in a digital medium, or in a mixture of the two.

We define collaboration as the activity of working together towards a common goal. There are a number of elements included in the definition. The first element is communication – the exchange of knowledge or opinions to achieve understanding by the partners. This element is a necessary but not sufficient condition for collaborative problem solving – it requires communication to go beyond mere exchange. The second element is co-operation, which is primarily agreed-upon division of labour. Co-operation in collaborative problem solving involves nuanced, responsive contributions to planning and problem analysis. An alternative view might regard co-operation as simply a lower-order version of collaboration, rather than as a component within it. A third element is responsiveness, implying active and insightful participation.

A contributing factor for effective collaborative problem solving is the accumulated expertise of the individuals concerned. The 21st-century media-rich environment stands in stark contrast to the face-to-face communication methods in earlier time periods. Communication structures based on technology inhibit the full range of human dynamics available in face-to-face environments. However, the skills needed to interpret human thought and behaviour within a problem-solving space are at least as formidable as those needed in face-to-face environments (Lee and Kim, 2005). Some aspects of face-to-face interactions may be relevant to the online environment, and others not. Factors specific to the online environment also need to be taken into account.

Figure 14.1 depicts the social and cognitive skills needed for collaborative problem solving. Collaboration requires “social interaction”, a term in need of an updated definition for the 21st century. For many, the terms “social” and “interaction” still imply sociability and face-to-face presence. This implication remains true for many situations. The identification of humans as social animals has a long history, and this identification has set humans apart from many other species. In more recent studies of human nature, the term “social” has also been applied to individuals’ preference to engage in activities that involve others or that are intended to help others, rather than sociability. In the context of collaborative problem solving, the term “social skills” refers to a set of processes that are concerned primarily with the style and quality of interaction, rather than with aspects of the task or problem. Dealing with the latter can be characterised as cognitive skill input, as learners work through problem-solving processes using content knowledge where relevant.

Figure 14.1. **Framework for collaborative problem solving**

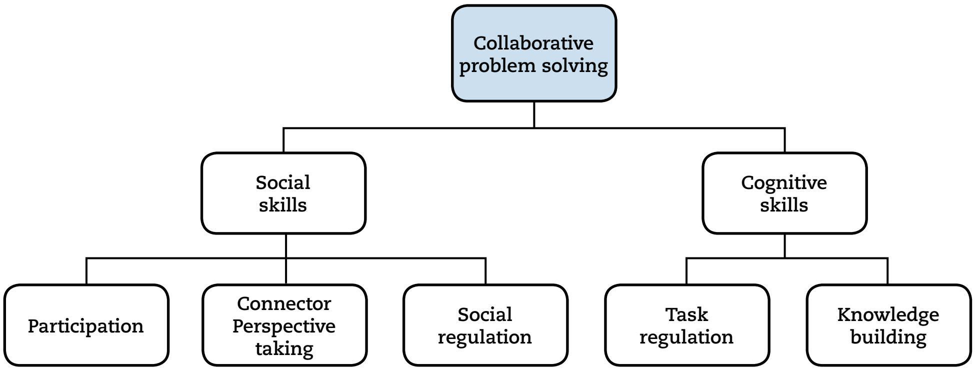


Figure 14.1 identifies strands within the two broad sets of skills. It shows the skills of participation, perspective taking, social regulation, task regulation, and knowledge building as strands broadly divided into social and cognitive skills, as described in the social sciences literature (see, for example, O’Neil, Chuang and Chung, 2003). Each of the five strands may be divided into more specific elements. For example, the skills and processes of negotiation, metamemory, transactive memory and responsibility all contribute to social regulation. Table 14.1 shows the conceptualisation of collaborative problem solving (Hesse et al., 2015) that underpins the following discussion of the design and development of assessment tasks and constitutes the theoretical framework within which results from student assessments are interpreted.

Table 14.1. **Components, strand elements and indicative behaviour needed for collaborative problem solving**

Social strands	
Elements	**Indicative behaviour**
Participation	
Action	Activity within environment
Interaction	Interacting with, prompting and responding to the contributions of others
Task completion	Undertaking and completing a task or part of a task individually
Perspective taking	
Adaptive responsiveness	Ignoring, accepting or adapting contributions of others
Audience awareness (mutual modelling)	Awareness of how to adapt behaviour to increase suitability for others
Social regulation	
Negotiation	Achieving a resolution or reaching compromise
Self evaluation (meta memory)	Recognising own strengths and weaknesses
Transactive memory	Recognising strengths and weaknesses of others
Responsibility initiative	Assuming responsibility for ensuring parts of the task are completed by the group

Cognitive strands	
Elements	**Indicative behaviour**
Task regulation	
Resource management	Managing resources or people to complete a task
Collect elements of information	Explores and understands elements of the task
Systematicity	Implements possible solutions to a problem and monitors progress
Tolerance of ambiguity/tension	Accepts ambiguous situations
Organisation (problem analysis)	Analyses and describes the problem in familiar language
Setting goals	Sets a clear goal for a task
Knowledge building	
Knowledge acquisition	Follows the path to gain knowledge
Relationships (representation and formulation)	Identifies patterns and connections among elements of knowledge associated with the problem
Rules “if…then”	Uses understanding of cause and effect to develop a plan
Hypothesis “what if …” (reflection on monitoring)	Adapts reasoning or course of action as information or circumstances change
Solution	Problem solved

Use of the collaborative problem-solving framework

This framework can be used to design and create tasks which will generate results that can be mapped onto developmental progressions. This framework takes the approach that all skills are both teachable and learnable. Although each of the strands is hypothesised to contribute to the overall construct, the framework does not specify how they relate to one another. Figure 14.1 is a simplistic graphic that does not take into consideration the degree to which the strands may be interdependent, or may contribute to each other. Similarly, different subskills will be required for different problems, depending on the problem characteristics. A general problem-solving process is described here in order to focus attention on those steps measured tasks designed to assess collaborative problem solving.

Processes involved in collaborative problem solving

As a first step, a problem needs to exist and to be identified. The process of identification itself can be a complex analytic endeavour. A straightforward approach is to ask “who”, “what”, “how” and “why”. This approach is supported by a meta-analysis of examples of collective intelligence (Malone, Laubacher and Dellarocas, 2007). The answers to these questions help to define the problem space; its current status, its goal and the artefacts available for manipulation within it.

The question “what is the problem?” helps to determine who ought to be involved and the approach that should be taken to problem solving. Problem spaces may be pre-existing (Malone, Laubacher and Dellarocas, 2007) or, as in the case of assessment contexts, be contrived. The next step is to answer the question “what is the goal?”. For example, is the goal an act of creation or of decision making? A subsequent issue is the “how” of collaborative problem solving. This question asks how the problem space should be regulated.

A collaborative problem-solving scenario posed early in the ATC21S project drew on a common school issue in the early 21st-century – how to deal with bullying. The sequence of steps identified above drew on the collaborative input of the group: What is the nature of the problem? What is the desired solution? Who is involved in the problem space? How will the learners collaborate within the space? All of these questions are influenced by answering the “why” question in this instance.

After identifying the problem (who, what and why), the next step involved deciding how (technology, environment) and when (timing, frequency, duration) to interact in the problem space. The second step involved agreeing on how to approach the problem (strategy), which required background background information about the problem and solutions to similar problems. The third step was to identify the tasks that needed to be completed and how they would be undertaken.

Each step was a miniature collaborative problem-solving cycle in its own right. The steps constitute cognitive processes, yet each required social skills to mediate and deliver them. Each step required participation, perspective taking, social regulation, task regulation, and knowledge building.

Learning progressions

Both specific and general development learning progressions describe the change in learners’ competence as they evolve from beginner to advanced. These progressions provide a basis for developing assessment tasks for learners at different skill levels. These tasks can also act as prototypes for teachers to use as learning tools in the classroom.

Developmental learning progressions exist. Rubrics were developed for each of the elements within the collaborative problem-solving framework in order to indicate different levels of skills for each element (see Griffin and Care, 2015). For example, exemplar performance quality criteria were created for task development for the “action” element within the “participation” strand (Table 14.1) and hence also for assessing learners' competencies. The development of the criteria was based on the literature, which supports the role of participation in the collaborative problem-solving process. The basic premise of the participation strand is that an effective collaborative problem solver has to be able to engage in the resolution of problem solving as a member of a group addressing a common goal. Assessing the action element of the participation strand, a student might need to meet the following quality criteria:

- is active independent of the group
- is active in the scaffolded environment
- activates and scaffolds others.

Collectively, the performance quality criteria developed for each element are expected to provide adequate and appropriate evidence that would satisfy an observer that a skill has been acquired and can be demonstrated. The detailed criteria describe development from less to more skilled behaviours. For each of the elements, several performance quality criteria were developed. These were nominally at "low", "middle" and "high" levels of operation.

Task creation took into consideration the collaborative-problem solving framework, and its detailed performance quality criteria. Figure 14.2 displays the stimulus and response flow of an exemplar assessment task. "Assessment task" refers to a linked set of stimuli that are sufficient to identify the existence of a problem with its corresponding artefacts. Within the assessment task, which could equally be labelled "problem space", a flow of items and coding follows, primarily through from Sub-task 1 for this illustration. Assessment tasks and their sub-tasks are not restricted to this number of tasks, prompts, responses or codes; the figure is illustrative only.

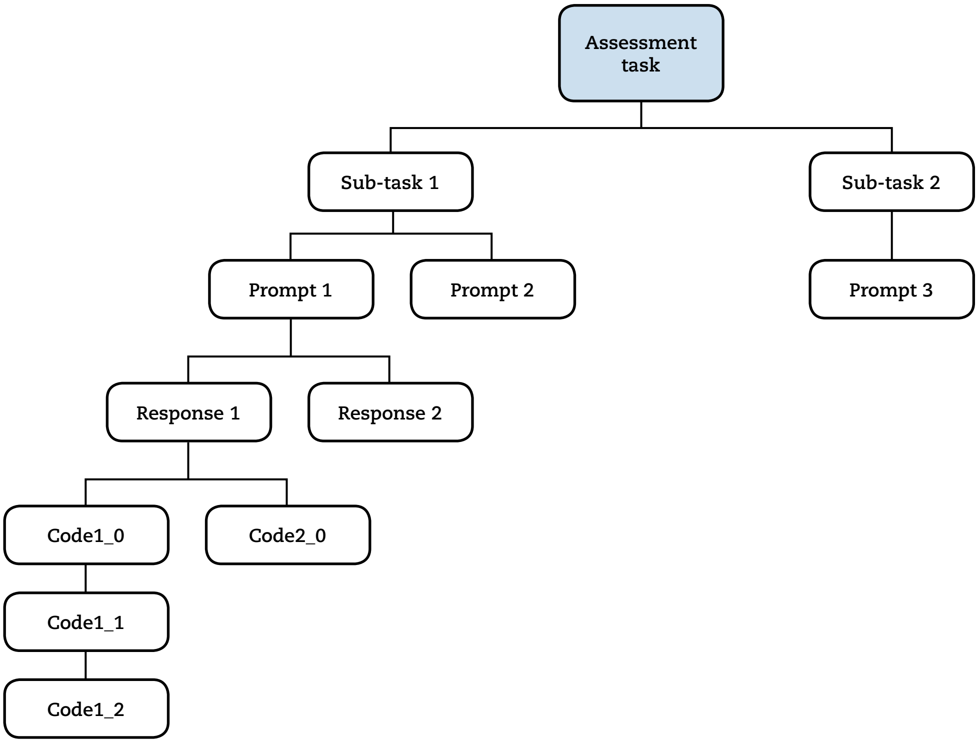
Figure 14.2 shows the problem space a learner enters in an electronic environment or scenario which presents a number of sub-tasks within it. The learner engages with the task in response to a number of prompts. The prompts vary in number and stimulate a number of possible responses. The responses may demonstrate different levels of quality of performance. The assessment and its sub-tasks are constructed so that multiple strands of the constructs of interest can be sampled. In Figure 14.2, for example, Sub-task 1 provides Prompts 1 and 2, which reflect particular capabilities, while Sub-task 2 provides Prompt 3, another capability. Due to practical constraints, it is unlikely that a single assessment task can provide learners with enough prompts to demonstrate their performance across all strands of the construct.

These constraints include the complexity of the collaborative problem-solving construct, the time available for assessment tasks and current limitations in computer programming. Accordingly, a set of assessment tasks can be created, each of which samples some of the skills needed for collaborative problem solving. Individual responses across the tasks generate a profile of variation in skills across the strands. This information can then be used by the teacher to target teaching to the student's zone of proximal development (Vygotsky, 1986).

Characteristics

Assessment tasks must be constructed to reflect the kind of problems that require collaboration. The characteristics of such problems are ambiguity, asymmetry and unique access to resources, with consequent dependence between learners. With such tasks, it is possible to test the construct definition model, the developmental learning progressions, the indicators of increasing competence, and the task development and delivery. At the simplest level, a problem-solving task can be designed that makes collaboration desirable or essential. In the classroom, this can be achieved by the teacher giving different sets of information to different students in a group, rather than giving them all the same information. In order to solve the problem, the students then need to collaborate in order to access the required resource – information. This classroom example mirrors real-life collaborative problem-solving situations, where information may be derived from different sources and is not shared beforehand.

Figure 14.2. **Illustration of an assessment task and of a stimulus-response-to-code structure**



Note: Sub-task 1 provides Prompts 1 and 2, which reflect particular capabilities, while Sub-task 2 provides Prompt 3, another capability. A single response can be differentiated into different codes, including sub-codes.

Giving learners unique access to different resources creates a dependence between them that provides a more authentic prompt for collaborative activity than mere instructions from a teacher for students to "work together". Working together may be valued for its social aspect, but might not be essential. Resources may reside in knowledge, tools or skills. Both the differential access to resources and the consequent dependence between students bring about asymmetry in the assessment task activity. Asymmetry raises some interesting challenges in the world of assessment. Most educational assessment presumes the individuals attempting the tasks face the same conditions and specifications. Accordingly, providing different sets of stimuli to the collaborating students immediately raises issues of the capacity of the tasks to provide comparable information across students.

Examples

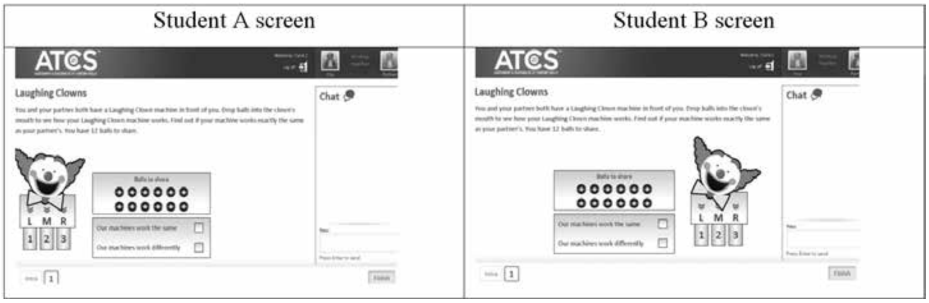
Two tasks are outlined here that can be completed by an individual who has access to all resources, or by a pair of students who are sharing resources. In the latter case, the students are dependent on each other to solve the problem. These simple tasks provide examples of how to re-structure a problem for collaborative work. Each task is presented initially with full information so one individual can solve the problem alone and then with collaborating students sharing the resources. The first task is structured symmetrically – both students have access to the same resources, and merely need to compare information in order to finalise the problem solution. The second task is structured asymmetrically – each student has access to different resources and neither is aware either of these differences or of the nature of the resources available.

A symmetric task

In this task (called "Laughing Clowns"), the student views a screen depicting a clown machine and 12 balls (Figure 14.3). The task is to work out how the machine functions, i.e., to determine from which slot the balls will appear. The student must place the balls into the clown's mouth while it rotates back and forth in order to determine the rule governing which slot the balls will emerge from. The student needs to:

1) collect information, that is, see that when balls are dropped, they come out of different slots
2) identify patterns, that is, whether the balls come out in a predictable way
3) form rules, for example, that balls placed when the clown has rotated to the left side will come out of a specific slot, when to the middle, another specific slot, and to the right, another slot
4) test rules, that is, whether the balls always follow the pattern.

Figure 14.3. **Symmetric task: "Laughing Clowns"**

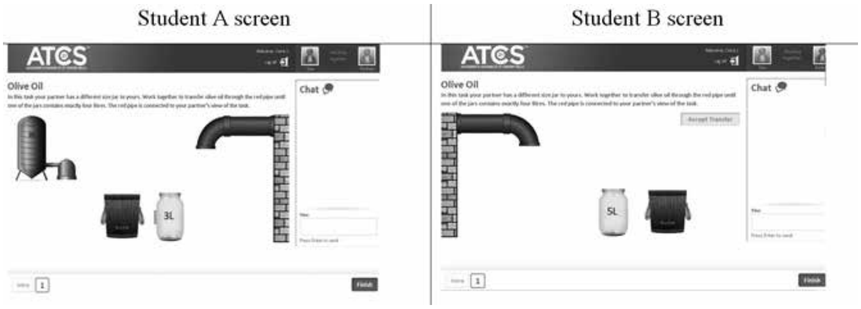


You and your partner both have a Laughing Clown machine in front of you. Drop balls into the clown's mouth to see how your Laughing Clown machine works. Find out if your machine works exactly the same as your partner's. You have 12 balls to share.

You and your partner both have a Laughing Clown machine in front of you. Drop balls into the clown's mouth to see how your Laughing Clown machine works. Find out if your machine works exactly the same as your partner's. You have 12 balls to share.

This task can be modified for collaborative problem solving and extended beyond collecting information, identifying patterns, and forming and testing rules. With multiple clown machines and multiple problem solvers, the problem can require cognitive problem-solving processes further up the hierarchy, such as generalising rules and generating and testing hypotheses. A simple expansion is to present a mirror image of the clown machine to each of two students (see Figure 14.3), who must share a set of balls, and instruct them to work out if the two machines work similarly. Although students are not told in advance that they have the same machines, they may infer this from the instruction that they are to identify if their machines work the same way. The core task remains the same – to identify patterns – but the students now have to communicate their results in such a way that they have sufficient evidence to compare. The resource – the balls – is also limited so students only have a finite number of steps in which to collect the information required. Practically, there are two independent steps – identification of pattern and comparison of results. These two steps are made more complex by the need to share resources and the need to communicate.

Figure 14.4. **Asymmetric task: "Olive Oil"**



In this task your partner has a different size jar to yours. Work together to transfer olive oil through the red pipe until one of the jars contains exactly four litres. The red pipe is connected to your partner's view of the task.

An asymmetric task

Another simple task is based on the style of the Tower of Hanoi problem popularised by mathematician Eduard Lucas in 1883 (Newell and Simon, 1972; Petkovi, 2009). This style of problem requires the solver to work out a sequence of movements to achieve the goal. It bears some resemblance to the forward planning requirements of a chess game – in thinking beyond one step to the next before implementing an action.

Although the Tower of Hanoi model represents a static problem, it can be made dynamic and less well-defined by not providing all the information required at the outset, and by distributing the resources. The problem solver needs to explore the environment in order to understand the resources available, and then work out the most efficient steps to effect a solution. In the version of the task presented called "Olive Oil", with two students collaborating, it becomes asymmetric and less well-defined. Figure 14.4 shows the different views presented to each student. Students must work out what artefacts and processes are needed to solve the problem. They need to establish how the artefacts function and whether they are necessary. In terms of processes, given the shared nature of the resources, they need to identify the series of steps to be undertaken. Having access to and knowledge of different artefacts redefines the problem, so that students need to solve the problem of task regulation before they can identify the sequence of movements they need to implement. This changes the nature of the problem, and incorporates tasks at more complex levels of the problem-solving hierarchy.

Assessing collaborative problem solving

According to Messick (1995) the main idea of construct validity is the credibility of the interpretation of assessment evidence. Therefore the evidential basis behind the interpretation of assessments becomes the substance defining the construct.

In ATC21S the evidence was interpreted in terms of the educational consequences for the teaching and learning of the construct – collaborative problem solving – and for the overall intended consequences for curriculum and policy determination (Darling-Hammond, 2012). The interpretation of student collaborative problem-solving behaviour in ATC21S:

- made it possible for teachers to understand more of the construct to teach and improve student performance in collaborative problem solving
- analysed and explored differences between individual student performances and the capacity to explain these using covariance analyses
- helped determine policy by aggregating data at a system level.

Human-to-human versus human-to-agent collaboration

The challenge for the ATC21S project was to design and deliver tasks for classroom use in real-life environments, where either face-to-face or e-collaboration could occur. Some have advocated an alternative approach (Franklin and Graesser, 1996; Blech and Funke, 2010). Instead of having two or more students interact (human-to-human, H2H) they replace one or more of the students in a digital context by an artefact such as an avatar or digital agent (human-to-agent, H2A). Collaboration or interaction could thus occur between:

- a digital agent and at least one human participant
- two or more digital agents with at least one human participant
- two or more agents without human participants.

This raises questions about how equivalent H2A interactions are to interactions between two or more people, but the strategy reflects the simplification used in many teaching interventions. Throughout their education, students often face problems that are not "authentic", but nonetheless constitute valuable learning experiences. Assessment often takes the same approach. Selected aspects of performance are assessed, and these are selected because they are hypothesised to be central to the domain of interest as well as to the viability of the exercise. In this instance, the question is whether the constraints upon individuals' problem-solving strategies brought about by the artificial structuring of the environment are severe enough to remove the assessment too far from the nature of the domain.

Messick (1995) required that both actual and potential consequences of interpretation within the applied setting should be studied. He recommended examining the advantages and disadvantages of the proposed assessment use against alternative approaches with the same intended purpose. If the assessment is properly designed, any undesirable consequences would not be the result of construct under-representation or construct-irrelevant variance. Construct under-representation occurs when essential dimensions of the construct are not incorporated in the assessment. To guard against this, a detailed definition and assessment framework must be developed and the assessment tasks constructed within this framework. Construct-irrelevant variance may derive from an external source of bias.

It is still to be determined whether the H2A approach is consistent with, or different from, a construct defined by the analysis of H2H interactions. There has been considerable research using H2A in collaborative learning (Campbell, 1968; Dillenbourg, 1999; Dillenbourg and Traum, 2006; Franklin and Graesser, 1996; Stahl et al., 2006) and it appears to be assumed that this transfers to collaborative problem solving. An elemental issue with both approaches concerns whether student responses are inhibited in the virtual environment with consequences for the underestimation of student skills.

Digital versus face-to-face interaction

Although ATC21S did not adopt the use of digital agents, it did take the approach of having two students interact via a digital medium. While this is considered relevant to 21st century interaction via the Internet, it is not as direct a measure as face-to-face interaction. The approach contributes to the discussion about the viability of measuring collaborative problem solving through H2A and H2H interaction via the Internet. It may be difficult to convince some stakeholders that virtual interaction, whether H2A or H2H, is collaboration in the true sense of the word (Funke, 1998, 2010). Both approaches may introduce construct-irrelevant variance as well as construct under-representation with consequential construct and face validity issues. In both approaches, a construct may well be defined and clearly delineated as a measurable entity, but these may not be the same constructs.

There are reasons to use either or both approaches. There will be differences in the representation of collaborative problem solving between events where persons interact face to face, and those where they interact through electronic media. However, both contexts allow for collaborative effort – two or more individuals working together toward a common goal. ATC21S was restricted to the use of two or more humans interacting in a digital context without face-to-face interaction (O'Neil, 1999; O'Neil, Chuang and Chung, 2003). Thus, in this case, construct under-representation was a potential problem. The ATC21S project took the view that H2H interaction would most closely approximate in-person collaborative problem-solving activity although the virtual context raises questions about authenticity.

The impact of asymmetric tasks

In ATC21S, attempts were made to design the assessment environment to be as close to the authentic context in which collaborative problem solving would take place as possible. For example, some tasks were designed to be asymmetric with respect to participant ability such that each participant had access to different sources of information (Adejumo, Duimering and Zhong, 2008). There was no systematic approach allocating students to roles – that is, which sets of resources they had control over or access to. This meant the loss of the potential for manipulation and control that can be afforded through the use of digital participants (Funke, 1998, 2010), but it is arguable whether this constitutes authentic collaborative behaviour, which is essentially uncontrolled.

One research interest of ATC21S was the use of aggregate data and the influence of the asymmetric allocation of student roles (A and B) on student and student-pair performance, and the consequences for general disability of interpretation of the collaborative problem-solving assessment data. One issue was the degree to which the capacity of one partner would affect the capacity of the other to demonstrate their skills. Another question centred on whether the fact that dyads were responding to asymmetric tasks would affect individual student scores and whether this would be reflected in estimates of population parameters. ATC21S only used dyads as undertaken by Schwartz (1995). There was no extension to larger numbers of collaborators.

Method

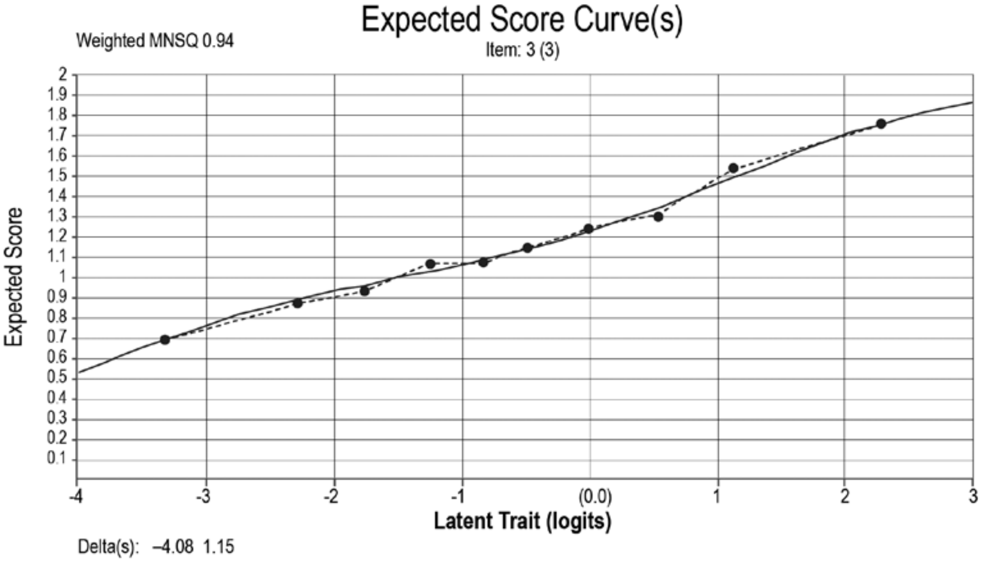
The measures were based on a range of approaches to coding, scoring and analysis, all focusing on the series of elements and their quality performance criteria which were developed and linked to the conceptual framework for collaborative problem solving described more fully in Hesse et al. (2015). The elements and quality performance criteria were used by human raters to analyse log files of student pairs across tasks. Log files included both process (or actions taken) streams and chat streams. A total of 1 552 ratings of student responses across 6 tasks was analysed. All raters coded both students for every pair they rated, generating data for 751 dyads.

Analysis

The project used one- and two-parameter item response models to explore the data and to determine its dimensionality. Two dimensions were hypothesised – the social and cognitive. Fit to the data was excellent with a one-parameter model. This was improved when the second parameter was used. An example of model fit is illustrated in the item characteristic curve presented in Figure 14.5. One parameter was preferred for interpretive purposes; two parameters were used for fit improvement.

Differential item function analysis was also conducted using Student A and Student B as the subgroups. It should be noted that students were only analysed within one dyad, limiting the conclusions that can be drawn from the analysis. Figure 14.6 illustrates an example of the differential item functioning analysis. All items presented similar results except the last item – problem solution. Even then, fit was very good and its differential item functioning was not relevant to the overall score.

Figure 14.5. **Example item characteristic curve observed (dotted line) and modelled (solid line)**



From the differential item functioning analysis it can be seen that it does not matter which student was Student A and which was Student B; that is, it does not matter which student has access to which resources. Note that this does not illustrate whether or not the asymmetry of abilities influences the performance of Student A or Student B, only that the different access to resources is not a determinant of the level of skill displayed.

The data were then split into two, and separate analyses conducted for Students A and Students B (Table 14.2). The correlation matrices generated a root mean square error (RMSE) of 0.02 after mapping the Student A matrix onto the Student B matrix. The near identical nature of the two correlation matrices provides more support for the interpretation that it matters little who was Student A and who was Student B within the dyads.

Figure 14.6. **Differential item functioning for Student A and Student B – Typical solution**

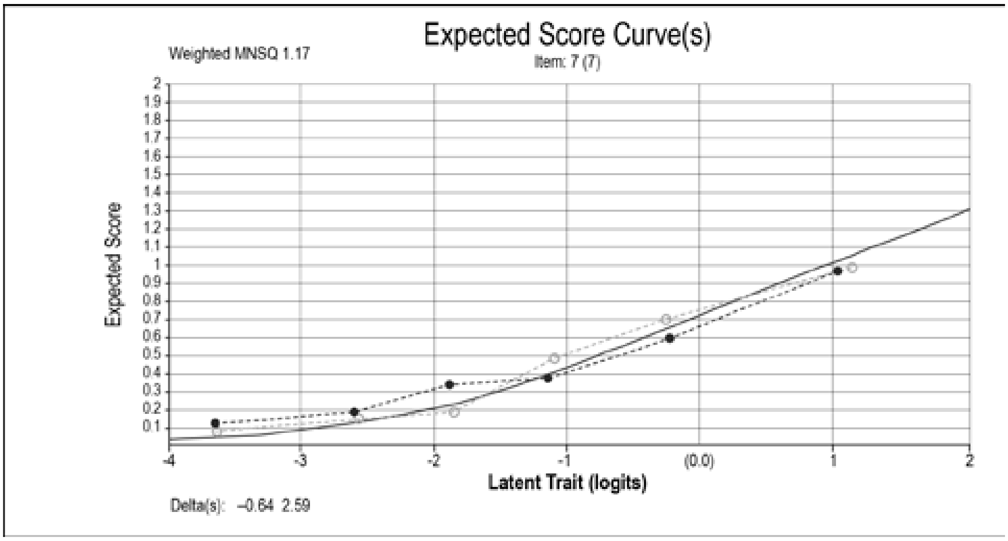


Table 14.2. **Correlation matrices for Student A and Student B across the social and cognitive strands and their components**

Student A	Social strand	Cognitive strand	Participation	Perspective taking	Social regulation	Task regulation
Cognitive strand	0.75					
Participation	0.87	0.67				
Perspective taking	0.88	0.69	0.67			
Social regulation	0.92	0.69	0.69	0.75		
Task regulation	0.72	0.95	0.64	0.68	0.66	
Knowledge building	0.7	0.95	0.63	0.63	0.64	0.81

Student B	Social strand	Cognitive strand	Participation	Perspective taking	Social regulation	Task regulation
Cognitive strand	0.73					
Participation	0.88	0.63				
Perspective taking	0.87	0.66	0.65			
Social regulation	0.91	0.68	0.68	0.72		
Task regulation	0.68	0.96	0.59	0.62	0.65	
Knowledge building	0.7	0.95	0.62	0.63	0.65	0.81

These two analytic approaches may contribute to the argument that, at an aggregate level, overall population parameters may not be subject to variations in ability within collaborative teams.

If the purpose of the H2A approach is to control one member of a collaborative team it is likely to be more successful than H2H. The question then to be addressed is the purpose of this control. If it is to control variation due to asymmetrical ability within the collaborative teams, these data go some way to suggest that that is not necessary. If it is to test specific hypotheses of how a particular dyad member would respond to specific stimuli within the collaborative problem, then H2A may well be the better approach. In this case, Messick's advice regarding construct-relevant variance would need to be considered. It should be noted that a lack of impact of access to different resources, and the consistency of correlations may be necessary conditions for establishing that variation in ability levels within a pair does not impact the partner's capacity to perform, but this is not sufficient. Variations in dyad pairs would need to be manipulated to reach such a conclusion.

Discussion

This chapter describes an approach to measuring collaborative problem solving as a 21st-century skill. We need to develop such skills in order to deal with global problems in a global environment. This approach can provide a guide to and stimulus for teachers to frame their pedagogical methods in ways that will model collaborative practices, problem solving, and learning and knowledge building for students.

In part because it is difficult to visualise what productivity in the future might mean, we need to implement education and training that will ensure optimal communication, effective learning and successful problem solving.

Problem solving is the key not only to enhancing productivity, but to doing it in a way that factors sustainability into the larger-scale dilemmas in our world. Drawing on perspectives from education (Vygotsky, 1986), anthropology and artificial intelligence (Lave and Wenger, 1991), and measurement (Griffin, 2009), we have outlined an example of a framework within which we can develop learning progressions amenable to assessment and teaching. In the short term, as society becomes more dependent on information-oriented skills, it needs to address different ways of working and thinking. Moreover, we will need a new range of tools for working and living. Education will continue to play a critical role in developing skills for the future. Pressure on individuals to participate in information-based problem resolution will increase. The global economy is changing, and this will generate structural change in the workplace and in the nature of work itself. As occupations based on the economic realities of the past are lost, new occupations will emerge in the production, distribution and consumption of information, rather than of commodities. In this world, the role of education in teaching and assessing new skills will become clearer. What happens in the classroom will be an integral aspect of preparing individuals and groups to take their place in this information-rich milieu. In the long term, those who can collaborate in their learning and problem-solving activities in a digital environment will be better placed to operate in the world beyond the information age. We have outlined an approach trialled through ATC21S within the classroom, but with applications beyond its walls. Student and teacher responses to the assessment tasks provide a focus for continuing work on this approach. ATC21S provides indications of the capacity to measure H2H collaborative problem-solving skills in a virtual environment. The next challenge lies in linking these measures to classroom behaviour, as a precursor to demonstrating the learnability of the skills and their application.

References

Adejumo, G., R.P. Duimering and Z. Zhong (2008), “A balance theory approach to group problem solving”, *Social Networks*, Vol. 30/1, pp. 83-99.

Autor, D., F. Levy and R.J. Murnane (2003), “The skill content of recent technological change: An empirical exploration”, *The Quarterly Journal of Economics*, Vol. 118/4, pp.1279-1333.

Binkley, M. et al. (2012), “Defining twenty-first century skills”, in P. Griffin, B. McGaw and E. Care (eds.), *Assessment and Teaching of 21st Century Skills,* Springer, Dordrecht, pp. 17-66.

Blech, C. and J. Funke (2010), “You cannot have your cake and eat it, too: How induced goal conflicts affect complex problem solving”, *Open Psychology Journal*, Vol. 3, pp. 42-53, http://dx.doi.org/10.2174/1874350101003010042..

Campbell, J. (1968), “Individual versus group problem solving in an industrial sample”, *Journal of Applied Psychology*, Vol. 52/3, pp. 205-210.

Darling-Hammond, L. (2012), “Policy frameworks for new assessments”, in P. Griffin, B. McGaw and E. Care (eds.), *Assessment and Teaching of 21st Century Skills,* Springer, Dordrecht, pp. 301-339.

Dillenbourg, P. (1999), “What do you mean by ‘collaborative learning’?” in P. Dillenbourg (ed.), *Collaborative-Learning: Cognitive and Computational Approaches*, Elsevier, Oxford, pp. 1-19, http://dx.doi.org/10.1207/s15327809jls1501.

Dillenbourg, P. and D. Traum (2006), “Sharing solutions: Persistence and grounding in multi-modal collaborative problem solving”, *Journal of the Learning Sciences*, Vol. 15/1, pp. 121-151.

Franklin, S. and A.C. Graesser (1996), "Is it an agent or just a program? A taxonomy for autonomous agents", *Proceedings of the Workshop on Intellingent Agents III, Agent Theories, Architectures, and Languages*, Springer, Berlin.

Funke, J. (2010), "Complex problem solving: A case for complex cognition?" *Cognitive Processing*, Vol. 11/12, pp. 133-142 , http://dx.doi.org/10.1007/s10339-009-0345-0.

Funke, J. (1998), "Computer-based testing and training with scenarios from complex problem-solving research: Advantages and disadvantages", *International Journal of Selection and Assessment*, Vol. 6/2, pp. 90-96, http://dx.doi.org/10.1111/1468-2389.00077.

Griffin, P. (2009), "Teachers' use of assessment data", in C. Wyatt-Smith and J. Cumming (eds.), *Educational Assessment in the 21st Century: Connecting Theory and Practice*, Springer, New York, pp. 183-208.

Griffin, P., E. Care and B. McGaw (2012), "The changing role of education and schools", in P. Griffin, B. McGaw and E. Care (eds.), *Assessment and Teaching of 21st Century Skills,* Springer, Dordrecht, pp. 1-15.

Griffin, P., B. McGaw and E. Care (eds.) (2012), *Assessment and Teaching of 21st Century Skills*, Springer, Dordrecht, http://doi.org/10.1007/978-94-017-9395-7.

Hesse, F., E. Care, J. Buder, K. Sassenberg and P. Griffin (2015), "A framework for teachable collaborative problem solving skills", in P. Griffin and E. Care (eds.), *Assessment and Teaching of 21 st Century Skills, Volume 2: Methods and Approach*, Springer, pp. 37-56, http://dx.doi.org/10.1007/978-94-017-9395-7_2.

Lave, J. and E. Wenger (1991), *Situated Learning: Legitimate Peripheral Participation*, Cambridge University Press, New York.

Lee, M. and D. Kim (2005), "The effects of the collaborative representation supporting tool on problem-solving processes and outcomes in web-based collaborative problem-based learning (PBL) environments", *Journal of Interactive Learning Research*, Vol. 16/3, pp. 273-293.

Malone, T.W., R Laubacher and C. Dellarocas (2007), "Harnessing crowds: Mapping the genome of collective intelligence", *Working Paper*, No. 2009–001, MIT Center for Collective Intelligence, Massachusetts Institute of Technology, Cambridge, MA.

Mayer, R.E. and M.C. Wittrock (1996), "Problem-solving transfer", in D.C. Berliner and R.C. Calfee (eds.), *Handbook of Educational Psychology*, Macmillan, New York, pp. 47-62.

Messick, S. (1995), "Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning", *American Psychologist*, Vol. 50/9, pp. 741-749.

Messick, S. (1989), "Validity", in R.L. Linn (ed.), *Educational Measurement*, 3rd edition, American Council on Education/Macmillan, New York, pp. 13-103.

Messick, S. (1975), "The standard problem: Meaning and values in measurement and evaluation", *American Psychologist*, Vol. 30/10, pp. 955-966.

National Research Council (2011), *Assessing 21st Century Skills: Summary of a Workshop,* National Academies Press, Washington, DC.

Newell, A. (1963), *A Guide to the General Problem-Solver Program GPS-2-2*, RAND Corporation, Santa Monica, CA.

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

O'Neil, H.F. (1999), "Perspectives on computer-based performance assessment of problem solving", *Computers in Human Behavior*, Vol. 15, pp. 255-268.

O'Neil, H. F., S. Chuang and G. Chung (2003), "Issues in the computer-based assessment of collaborative problem solving", *Assessment in Education*, Vol. 10/3, pp. 361-373.

Petkovi, M. (2009), *Famous Puzzles of Great Mathematicians*, American Mathematical Society, Providence, RI.

Schwartz, D.L. (1995), "The emergence of abstract representations in dyad problem solving", *Journal of the Learning Sciences*, Vol. 4/3, pp. 321-354.

Stahl, G., Koschmann, T. and Suthers, D. (2006), "Computer-supported collaborative learning: An historical perspective", in R.K. Sawyer (ed.), *Cambridge Handbook of the Learning Sciences,* Cambridge University Press, Cambridge, UK, pp. 409–426.

Vygotsky, L.S. (1986), *Thought and Language*, The MIT Press, Cambridge, MA.

Vygotsky, L. S. (1978), *Mind and Society: The Development of Higher Mental Processes*, Harvard University Press, Cambridge, MA.

Chapter 15

Assessing conversation quality, reasoning, and problem solving with computer agents

By
Arthur C. Graesser
Psychology Department, University of Memphis, United States.
Carol M. Forsyth
Educational Testing Service, San Francisco, United States.
and Peter Foltz
University of Colorado at Boulder and Pearson, United States.

Conversational agents have been developed to help students learn in computer-learning environments with collaborative reasoning and problem solving. Conversational agents were used in the 2015 Programme for International Student Assessment (PISA) collaborative problem-solving assessments, where a human interacted with one, two or three agents. This chapter reviews advances in conversational agents and how they can help students learn by engaging in collaborative reasoning and problem solving. Using the example of AutoTutor it demonstrates how dialogues can mimic the approaches of expert human tutors. Conversations with intelligent systems are quite different depending on the number of agents involved. A human interacting with only one computer agent during a dialogue needs to continuously participate in the exchange. In trialogues there are two agents, so there are more options available to the human (including social loafing and vicarious observation) and the conversation patterns can be more complex, illustrated by Operation ARIES!, which uses a number of patterns in teaching students the basics of research methodology. Conversational agent systems use online, continuous, formative assessment of human abilities, achievements, and psychological states, tracked during the course of the conversations. Some of these formative assessment approaches are incorporated in the PISA 2015 assessment of collaborative problem solving. However, this chapter focuses on formative assessment in learning environments rather than on summative assessments.

Introduction

The OECD chose collaborative problem solving (CoPS) as a new assessment for the 2015 Programme for International Student Assessment (PISA) survey of student skills and knowledge (Graesser, Foltz et al., in press; OECD, 2013). The choice of CoPS reflects the greater focus in education on preparing students for college and careers by developing effective problem-solving skills and the ability to work with others to solve those problems.

The problem-solving dimension of the CoPS 2015 framework incorporated the PISA 2012 problem-solving framework targeting individual problem solving (Funke, 2010; OECD, 2010). As discussed in many of the chapters in this volume, the four cognitive processes of the individual problem-solving assessment were: 1) exploring and understanding; 2) representing and formulating; 3) planning and executing; and 4) monitoring and reflecting. The collaboration dimension of CoPS 2015 identified three proficiencies: 1) establishing and maintaining shared understanding; 2) taking appropriate action; and 3) establishing and maintaining team organisation. Multiplying these 4 problem-solving processes with the 3 collaboration proficiencies creates a matrix of 12 skills that represent the competencies of CoPS. A satisfactory assessment of CoPS would assess the skill levels of students for each of the 12 cells in this matrix. These skill levels then contribute to a student's overall CoPS competency score.

One heavily discussed issue in the development of the PISA CoPS 2015 assessment was the notion of an agent. The PISA CoPS 2015 framework used the following definition of CoPS (OECD, 2013):

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.

In this definition, the student would interact with either a human or computer agent. Indeed, students interacted with one, two or three computer agents on the assessment based on the 2015 framework. Therefore, the overall competency measure assessed how well a single human interacted with computer agents during the course of problem solving. The unit of analysis for the competency was the individual 15-year-old human interacting in a group, rather than the group as a whole.

There has been considerable discussion about having the students interact with computer agents rather than other humans in the CoPS 2015 assessment (see Graesser, Foltz, et al., in press; Rosen and Foltz, 2014). There were worries that the interaction with computer agents would not be sufficiently similar to interactions with fellow humans. Studies underway analysing the similarities and differences between students' interactions with computer agents versus other humans during problem solving (Greiff, personal communication). Nevertheless, logistical constraints led to the decision to have students interact with computer agents. The constraints of PISA required a computer-based assessment to measure students' CoPS skills in a short period of time (two 30-minute sessions) with two to three problem-solving scenarios per session. In such an assessment situation, the computer agents could presumably provide more control over the interaction and could provide sufficiently valid assessments of the constructs of interest within the time constraints. The system could measure a student's CoPS competency in multiple teams, with multiple tasks and multiple phases in a controlled interaction. This would be logistically impossible to do using human-to-human interaction. It often takes five minutes or more for a new group of humans to get acquainted in computer-mediated conversation before any of the actual problem-solving processes begin (e.g. Dillenbourg and Traum, 2006; Fiore and Schooler, 2004). Computer agents provide control over the logistical problems stemming from 1) assembling groups of humans (via computer mediated communication) in a timely manner within school systems with rigid time constraints; 2) the need to have multiple teams per student to obtain reliable assessments; and 3) measurement uncertainty and error arising when a student is paired with humans who do not collaborate. A systematic design of computer agents was ultimately created that provided control, many activities and interactions per session, and multiple groups of agents.

Questions continue to be raised about the status of the computer agents in assessments of collaborative problem solving. Many of the questions reflect uncertainties about the capabilities of computer agents at this point in their development. However, there have been recent advances in computer science, artificial intelligence, automated natural language understanding, computational linguistics, discourse processing, intelligent tutoring systems, informatics, text-to-speech generation, avatar design and other areas that feed into agent technologies. This chapter describes some of the learning environments where computer agents interact with humans in collaborative reasoning and problem solving. The hope is that readers acquire a more concrete understanding of what can be achieved in these environments with computer agents.

There is one caveat to note about this chapter. This chapter focuses on formative assessment in learning environments rather than summative assessment. Therefore, we will not elaborate on the specific assessment mechanisms in PISA CoPS 2015 or any other large-scale summative assessment. Nevertheless, the underlying approaches used in formative assessments are very relevant to the summative assessments and thus show the capability of systems with computer agents. Also, most of the learning environments discussed in this chapter focus on deeper levels of reasoning and problem solving, as opposed to the basic skills of literacy, numeracy, memory, perception and scripted procedures.

This chapter has three sections. The first section describes what computer agents are and identifies recent learning environments with conversational agents. The subsequent section describes a dialogue system called AutoTutor. AutoTutor helps students learn by interacting with them in natural language as they collaboratively solve problems or answer difficult questions that require reasoning. The last section describes systems with trialogues (two agents interacting with the human). The added agent provides additional capabilities that could not be provided by a single agent alone. In all of these systems, the computer continuously tracks the knowledge, skills, actions and psychological states online during the conversational interaction and logs these events so that the discourse moves on and actions of the computer agents are intelligently adapted.

Learning environments with conversational agents

Researchers have been developing learning environments with conversational agents for over two decades. Some of these environments attempt to help students learn by holding a conversation in natural language during the process of solving a problem or answering a deep-reasoning question. For example, AutoTutor (Graesser, 2016; Graesser et al., 2015; Graesser, Li and Forsyth, 2014; Graesser, Lu et al., 2004; Kopp, Britt, Millis and Graesser, 2012; Nye, Graesser and Hu, 2014; VanLehn et al., 2007) is a system that helps college students learn about computer literacy, physics, comprehension strategies and scientific reasoning. The agent is a talking head that speaks, points, gestures and exhibits facial expressions. One version of AutoTutor tracks the learners' emotions (such as confusion, frustration, and boredom/disengagement) through multiple channels of communication and adaptively responds with dialogue moves that are sensitive to the learner's affective and cognitive states (D'Mello and Graesser, 2012).

There are many other examples of agent-based environments that have successfully improved student learning through conversational mechanisms requiring reasoning and/or problem solving. Some focus on science, technology, engineering and mathematics (STEM) areas, (DeepTutor, Rus et al., 2013 and My Science Tutor, Ward et al., 2013; on biology (Betty's Brain, Biswas et al., 2010; GuruTutor, Olney et al., 2012; and Crystal Island, Rowe et al., 2010), or on basic electricity and electronics (BEETLE II, Dzikovska, et al., 2014), computer programming (Coach Mike, Lane et al., 2011) or mathematics (ALEKS, Nye, Graesser and Hu, 2014). Agent-based learning environments have been developed for various skills and strategies that require reasoning, such as asking and answering deep questions (iDRIVE, Gholson et al., 2009), metacognition and self-regulated learning (MetaTutor, Azevedo et al., 2010), comprehension (iSTART, Jackson and McNamara, 2013; McNamara et al., 2006), and research methods (Operation ARIES and ARA, Forsyth et al., 2014; Halpern et al., 2012; Millis et al., 2011). These learning environments have shown improvements in learning over and above comparison conditions.

These systems attempt to understand the student's natural language during the conversational interaction. This is possible because of advances in computational linguistics (Jurafsky and Martin, 2008; McCarthy and Boonthum-Denecke, 2012) and statistical representations of world knowledge and discourse meaning (Foltz, Kintsch and Landauer, 1998; Landauer, Landauer et al., 2007; Foltz and Laham, 1998). These advances automated analyses of word decomposition, sentence syntax, speech act classification, sentence semantics, text cohesion, and other aspects of language and discourse. Some of these advances are described later in this chapter.

Nevertheless, these agent-based systems do not always need to understand natural language. An alternative option is for the humans to select a verbal response or action from a menu of alternatives at each point in the conversation when they are expected to respond. For example, a recent project at the Center for the Study of Adult Literacy developed systems with agents that limited the human to clicking on options during the interaction (Graesser et al., 2015). These struggling adult readers had little or no ability to generate verbal responses so the input options were limited to click, scroll, drag, and drop. This allowed students to be trained and assessed in comprehension without being overly tasked on production. Moreover, automated analysis of natural language can be logistically impractical in an assessment involving several different languages. For example, within PISA CoPS 2015, communication between the human and the agents(s) was performed via a chat facility which typically presented four communication options for the human to select at each turn in a chat interaction. The options were judiciously generated by the item developers to assess the particular skills that are part of collaborative problem solving competencies (i.e., the particular cells in the "12-cell" matrix discussed earlier). Human physical actions, for example clicks on a diagram, were also limited. This allowed item developers to apply standard item response theory and other psychometric methods that have been developed over the decades for multiple choice tests.

A broad spectrum of computer agents has been used in tasks that involve tutoring, collaborative learning, co-construction of knowledge and collaborative problem solving. At one extreme are fully embodied conversational agents in a virtual environment with speech recognition, such as the Tactical Language and Culture System (Johnson and Valente, 2008) and assessment environments currently being developed at Educational Testing Service for English language learning, science and mathematics (Zapata-Rivera et al., 2015). Most systems have two-dimensional agents created from commercially available kits with text-to-speech engines and with the agent persona rendered with alternative facial features, clothing, facial expressions, gaze directions and gestures. Some systems instead use static pictures of agents with messages either spoken or presented in a print format to be read. A minimalist agent would have an icon persona and a printed text message. The PISA CoPS 2015 items had minimalist agents that consisted of printed messages in windows on the computer display, such as email messages, chat messages and documents in various social communication media. These forms of agent-based social communication media were also implemented in the assessment of problem solving in the Programme for the International Assessment of Adult Competencies (PIAAC; PIAAC Expert Group in Problem Solving in Technology-Rich Environments, 2009). It was very important to have the minimalist agents in these international assessments in order to minimise biases from alternative languages and cultures.

An assessment requires the human to pay attention to the agent when the agent communicates, just as when a human takes the floor when speaking. This can also be accomplished with a minimalist agent in a chat environment by the dynamic highlighting by dynamic highlighting of messages and windows through use of colour or flashing, and by co-ordinating messages with sounds (Mayer, 2009).

One major challenge in agent-based environments is managing the conversation. The system needs to interpret what the human says or does and to have a sensible response to each interpretation. Researchers create state-transition networks (Morgan et al., 2012; Person et al., 2001), chat maps (Zapata-Rivera et al., 2015), authoring tools with visualisation (Cai, Graesser and Hu, 2015), or other conversation flow tools to track the conversation options and make sure the system coverage is complete. This tracking process is most straightforward when the human can only choose from a small number of alternative responses. It gets tricky when there is natural language input from the human. For example, in a mixed-initiative dialogue system – a system that allows both the human and the artificial agents to direct the conversation – humans sometimes ask questions or change the topic. The computer system cannot always provide helpful or relevant responses to these agenda-changing dialogue moves. The computer needs to dismiss and redirect the conversation in these cases, for example with, “Terrific question but I don’t have an immediate answer.”, “How would you answer your question?”, or “Let’s stay on the main topic”. These difficult conversation junctures do not occur in agent-based systems that always limit the human’s input options. Some have worried that this limitation makes the conversation somewhat artificial compared with human-to-human interactions. Fortunately, this limitation is not serious in intelligent tutoring systems because the tutor sets the agenda in such systems, student questions and topic changes are comparatively infrequent, and students give up trying to ask questions and change the topic after they discover the responses are not that helpful (Graesser, McNamara and VanLehn, 2005).

Having described the properties and capabilities of agents used in formative and summative assessment, we now turn to some specific examples of agent environments that help students learn. The next section describes AutoTutor dialogues between a human and a single tutor agent, while the subsequent section describes systems with a second agent in conversational trialogues.

Conversational dialogues with AutoTutor

AutoTutor (Graesser, 2016; Graesser, Jeon and Dufty, 2008; Graesser, Lu et al., 2004; Nye, Graesser and Hu, 2014; VanLehn et al., 2007) is an intelligent tutoring system in which one human student and a single computer agent discuss topics such as computer literacy, physics, electronics and other STEM topics. AutoTutor is successful at tutoring students, achieving learning gains with effect sizes between $\sigma = 0.20$ and $\sigma = 1.00$ (Graesser, D’Mello and Cade, 2011; Nye et al., 2014; VanLehn, 2011) compared to attending classroom lessons and reading textbooks for an equivalent amount of time. The learning gains are comparable to those achieved by expert human tutors.

The structure of tutorial dialogue

The conversational structure of AutoTutor is based on fine-grained systematic qualitative and quantitative analyses of human-to-human expert tutoring sessions (Graesser and Person, 1994; Graesser, Person and Magliano, 1995; Person and Graesser, 1999). AutoTutor simulates some common conversation patterns in human tutoring. The most prominent are described below.

Tutoring sessions are organised around problems, challenging questions and tasks

The outer loop of tutoring consists of the selection of major tasks for the tutor and tutee to work on (VanLehn, 2006) whereas the inner loop consists of the steps and dialogue interactions to manage the interaction within these major tasks. After the tutor and tutee settle on a major task (outer loop), the tutor guides the specific agenda within the task (inner loop). In this sense, the typical tutoring interaction is tutor centered, not student centered.

A 5-step tutoring frame guides the major task (outer loop)

Once a problem or difficult main question is selected to work on, AutoTutor launches a 5-step tutoring frame (Graesser and Person, 1994):

1) Tutor asks a difficult question or presents a problem.

2) Student gives an initial answer.

3) Tutor gives short feedback on the quality of the answer.

4) Tutor and student have a multi-turn collaborative dialogue to improve the answer.

5) Tutor assesses whether the student understands the correct answer.

Step 4 in this 5-step tutoring frame involves collaborative discussion and joint action, with encouragement for the student to contribute knowledge rather than merely receiving knowledge. Step 4 is the heart of the collaborative interaction.

Expectation and misconception tailored dialogue guides micro-adaptation (the inner loop)

Human tutors typically have a list of expectations (anticipated good answers, steps in a procedure) and a list of anticipated misconceptions (incorrect beliefs, errors, bugs) associated with each task. For example, expectation E1 and misconception M1 are relevant to the example physics problem below:

PHYSICS QUESTION: If a lightweight car and a massive truck have a head-on collision, upon which vehicle is the impact force greater? Which vehicle undergoes the greater change in its motion, and why?

E1: The magnitudes of the forces exerted by A and B on each other are equal.

M1: A lighter/smaller object exerts no force on a heavier/larger object.

The tutor guides the student in articulating the expectations through a number of dialogue moves: pumps, hints and prompts for the student to fill in missing words. A “pump” is a generic expression to get the student to provide more information, such as “What else” or “Tell me more”. Hints and prompts are used by the tutor to get the student to articulate missing content words, phrases, and propositions. A hint tries to get the student to express a complex idea (as a proposition, clause or sentence) whereas a prompt is a question that tries to get the student to express a single word or phrase. For example, a hint to get the student to articulate expectation E1 might be “What about the forces exerted by the vehicles on each other?” This hint would ideally elicit the answer “The magnitudes of the forces are equal”. A prompt to get the student to say “equal” would be “What are the magnitudes of the forces of the two vehicles on each other?” The tutor generates an assertion if the student fails to express the expectation after multiple hints and problems. As the student and tutor express information over many turns, the list of expectations is eventually covered and the main task is completed. The tutor also has the goal of correcting misconceptions that arise in the student’s responses. When the student articulates a misconception, the tutor typically acknowledges the error and corrects it.

Tutor turns are well structured

During step 4 of the 5-step frame, most of the tutor’s turns have three informational components:

> Tutor turn: Short feedback + dialogue advancer + floor shift. The first component is short feedback (positive, neutral or negative) on the quality of the student’s last turn. The second component is a dialogue advancer that moves the tutoring agenda forward with either pumps, hints or prompts; assertions with correct information; corrections of misconceptions; or

answers to student questions. The third component shifts who has the conversational "floor" with cues from the tutor to the student. For example, a human tutor ends each turn with a question or a gesture to cue the student to do the talking.

Assessments of conversation and reasoning in the inner loop

Expectation and misconception tailored (EMT) dialogue provides the foundation for assessing student reasoning and problem solving in both AutoTutor and allegedly human tutoring. The basic premise is that for each task (i.e. the main question asked or problem to be solved) there are a set of expected answers (typically one to seven sentences or symbolic expressions) that are fragments of the content that would be part of the reasoning or solution to the problem. As the collaborative conversation proceeds, pattern recognition mechanisms are used to match the verbal behaviour and actions of the student to the expected answers. Match scores that meet or exceed some threshold for a particular expectation indicate that the student has successfully covered the expectation. When a match score falls below the threshold, the student has not covered the expectation, so the tutor needs to present scaffolding dialogue moves (pumps, hints or prompts) in an attempt to achieve successful pattern completion. For example, suppose an expectation has constituents A, B, C and D and the student has expressed A and B, but not C and D. The tutor would generate hints and prompts to attempt to get the student to cover C and D to meet the threshold for the expectation. The tutor gives up trying to get the student to cover the expectation after a few hints and prompts, ultimately generating an assertion that covers the expectation in the conversation. After all of the expectations for the task are covered, the task is finished and the tutor often gives a summary.

The assessment of the student's performance on a task can be calculated in a number of ways. One way is to compute initial match scores for each expectation immediately after the student gives their first response to a task. That is, there is an initial match score for each of the set of *n* expectations (E_1, E_2, … E_n) and an overall mean initial match score over the set of expectations. A second way is to assess performance is to compute students' cumulative match scores after they have received all of the scaffolding moves of AutoTutor. A third way is to systematically generate the scaffolding dialogue moves for each expectation and then compute how much scaffolding was attempted. For example, many versions of AutoTutor provide cycles of pump → hint → prompt → assertion for each expectation after the student's initial response to the main task. The scoring is specified below.

1. If the threshold is met for an expectation E_i after the main task, the student gets a full 1.00 credit for expectation E_i. The system moves on to cover another expectation.
2. If the threshold is not met in step 1, then the system delivers a pump and a match score is computed for the student's cumulative contributions for this task up to the student contributions after the prompt. If the match score meets the threshold after the pump, the student receives a score of 0.75 for expectation E_i and the tutor moves on to another expectation.
3. If the threshold is not met in step, AutoTutor generates a hint to elicit the uncovered constituents of expectation E_i. If the match score meets the threshold after the hint, the student receives a score of 0.50 for expectation E_i and the tutor moves on to another expectation.
4. If the threshold is not met in step 3, AutoTutor generates a prompt to elicit an uncovered constituent of expectation E_i. If the match score meets the threshold after the prompt, the student receives a score of 0.25 for expectation E_i and the tutor moves on to another expectation.
5. If the threshold is not met in step 4, AutoTutor generates an assertion, the student receives a score of 0 for expectation E_i, and the tutor moves on to another expectation.

According to this assessment algorithm, students score 0, .25, .50, .75, or 1.00 for each expectation and their overall score would be an average of the n expectations.

AutoTutor's assessment scores of course depend on how the match scores are computed. We have evaluated many semantic matchers over the years. The best results come from a combination of latent semantic analysis (LSA; Foltz, Kintsch and Landuaer., 1998; Landauer et al., 2007), frequency-weighted word overlap (rarer words and negations have higher weight) and regular expressions (which are beyond the scope of this chapter to define, see Jurafsky and Martin, 2008). In fact, LSA plus regular expressions have had high agreement scores in direct comparisons with human experts, compared with pairs of human experts (Cai et al., 2011). Interestingly, syntactic parsers did not prove useful in these analyses because a high percentage of the students' contributions are telegraphic, elliptical and ungrammatical. The underlying technique has been applied successfully to assessing team dialogues in group collaborative situations (Foltz, Laham and Derr, 2003; Foltz and Martin, 2008; Gorman et al., 2003; Martin and Foltz, 2004) suggesting it is appropriate to use it as a basis for computer agents for collaboration.

It is important to point out what AutoTutor does not compute in its assessment of reasoning and problem solving. AutoTutor does not directly track/measure deductive, inductive and other types of logical reasoning unless the expectations, hints and prompts are cleverly composed and ordered to focus on particular steps with associated content. AutoTutor does not directly track/measure particular steps, procedures and phases of problem solving unless the expectations, hints, and prompts are set up with these processing components having distinctive content. Instead, AutoTutor tracks the content through pattern-matching processes and delivers dialogue moves in attempts to achieve pattern completion of the content.

Trialogues

Trialogues are three-party conversations that extend the conversational framework of AutoTutor by incorporating two computer agents (Graesser, Forsyth and Lehman, in press; Graesser, Li and Forsyth, 2014). In addition to AutoTutor, many learning environments have incorporated multiple agents, such as Betty's Brain (Biswas et al., 2010), iSTART (Jackson and McNamara, 2013) and Operation ARIES! (Forsyth et al., 2013; Halpern et al., 2012; Millis et al., 2011).

Trialogues provide additional tutoring and assessment capabilities over and above the AutoTutor dialogues. In essence, sometimes two heads (computer agents) are better than one Graesser, Forsyth and Lehman, in press. That being said, there have been no direct comparisons in learning between the dialogues and trialogues with agents. However, researchers have recently been exploring several configurations of trialogues in order to better understand how they can be productively implemented for particular learners, subject matter, depths of learning and assessments (Cai et al., 2014; Graesser, Forsyth and Lehman, in press; Graesser, Li and Forsyth, 2014; Lehman et al., 2013; Zapata-Rivera, Jackson and Katz, 2015). For example, consider the trialogue designs below, which are relevant to the assessment of collaborative problem solving in addition to learning.

1) **Vicarious observation with limited human participation**. Two agents communicate with each other and exhibit social interaction, answers to questions, problem solving, or reasoning. The two agents can be peers, experts or a mixture. The agents can periodically ask the human a prompt question (inviting a yes/no or single word answer) that can be easily assessed automatically.

2) **Human interacts with two peer agents of varying proficiency.** The peer agents can vary in knowledge and skills. In assessment contexts, the computer can track whether the human is responsive to the peer agents, correctly answers peer questions, corrects an incorrect contribution by a peer, and takes initiative in guiding the exchange.

3) **Expert agent staging a competition between the human and a peer agent**. The human and peer agent play a competitive game (scoring points), with the competition guided by the expert agent.

4) **Expert agent interacting with the human and peer agent**. There is an exchange between the expert agent and the human, but the peer agent periodically contributes and receives feedback. Negative short feedback can be given to the peer agent on bad answers (the agent takes the heat) that are similar to the human's contributions.

5) **Human teaches/helps a peer agent with facilitation from the expert agent**. This is a "teachable agent" design (learning by teaching), with the expert agent rescuing problematic interactions.

6) **Human interacts with two agents expressing false information, contradictions, arguments or different views**. The discrepancies between agents stimulate cognitive disequilibrium, confusion and potentially deeper learning. This can dig deeper into subtle distinctions that often are important for assessment.

The remainder of this section describes an application of trialogues that was incorporated in a serious game to help students learn scientific reasoning. The system is called Operation ARIES! (Forsyth et al., 2013; Millis et al., 2011); Operation ARA was a similar web version in a beta release by Pearson Education (Halpern et al., 2012). ARIES/ARA (hereafter referred to as ARIES) has produced learning gains of approximately 1 sigma in a pre-test → interaction → post-test design. The game-like features include a storyline that extra-terrestrial invaders have conquered the world and propagated it with bad science. The human students must learn research methodology to defeat the invaders and save the world. The tutorial interactions during problem solving about research methodology occur across three distinct modules of ARIES: Cadet Training, Proving Ground and Active Duty. The three modules teach students the basic factual information about research methods, require them to apply this knowledge to specific cases, and train them how to ask good questions that diagnose whether a scientist is an alien peddling bad research. The system tracks the student's performance in mastering 11 core concepts with associated potential flaws about research methodology, such as premature generalisation of results, lack of a control group, and mistaking correlation for causation. We describe the three modules and some findings about learning, affect and assessment in the game.

Trialogues during factual learning in the Cadet Training module

In the Cadet Training Module, students learn the basic didactic information about the 11 core concepts through an e-text, multiple-choice questions and natural language trialogues between a human student and two artificial agents (a teacher agent and a student agent). The students are asked to answer multiple-choice questions because these questions serve as an assessment of prior-knowledge and aid learning through repeated testing (Roediger, McDaniel and McDermott, 2006). After completing the questions, students have an opportunity to read the e-text to obtain more information about the core concepts. The students are then tested again with multiple-choice questions and proceed to have adaptive conversations with the teacher agent, "Dr. Quinn", and the student agent, "Glass", based on their performance in the multiple-choice questions during training. The testing makes it possible to gauge students' knowledge levels (high, medium or low) on particular core concepts. Based on their assessed knowledge level, students then participate in one of three potential types of trialogues: vicarious learning, normal tutoring and teachable agent.

Students with low levels of knowledge in a core concept take part in vicarious learning: they simply watch the teacher agent teach the student agent with minimal interaction with of their own (e.g. simply answering yes/no questions). The purpose of asking the students these questions is to maintain their engagement and assess their performance. Previous research suggests that students with low knowledge learn best from vicarious learning because they have limited ability to generate relevant accurate content (Chi, Roy and Hausmann, 2008; Craig et al., 2012; Driscoll et al., 2003).

Tutorial interactivity is more conducive to learning for students with higher levels of knowledge (Jackson and Graesser, 2006; VanLehn et al., 2007). Therefore, students with an intermediate level of prior knowledge interacted through normal tutoring with the short feedback and the pump → hint → prompt → assertion scaffolding described above. The peer student periodically contributes to the conversation. Students with the highest level of knowledge teach the virtual artificial agent with a conversational pattern that is similar to a teachable agent mode (Biswas et al., 2010; Leelawong and Biswas, 2008; Roscoe, Wagster and Biswas, 2008).

Halpern and colleagues (2012) compared the three different types of trialogues (vicarious learning, normal tutoring and teachable agent) and a control (no conversation) in a pre-test → interaction → post-test design with an initial post-test and a second delayed post-test. There were no statistically significant differences in learning on the initial post-test between the three trialogue modes regardless of prior knowledge level. However, any of the three different types of trialogues were better than the control condition. Moreover, the teachable agent condition had greater learning gains on a delayed post-test than the vicarious learning condition. It is important to keep in mind that these three trialogue conversation modes were investigated in the context of didactic training only. Differences between the three modes may perhaps be found in learning environments requiring students to apply their knowledge rather than simply memorise facts as in the Cadet Training module.

Trialogues for case-based reasoning in the Proving Ground module

Different types of conversations may be more appropriate for learning environments where students are required to apply critical thinking strategies during case-based reasoning, as opposed to learning didactic information. The conversations in the Proving Ground module of ARIES are quite different from those in the Cadet Training module. The Proving Ground conversations require students to critically evaluate research cases, apply their didactic knowledge in their reasoning and achieve important deep-level discriminations that are known to correlate with learning (Anderson et al., 1995; Forsyth et al., 2013; VanLehn et al., 2007). The conversations embedded in the Proving Ground module required active reasoning in the conversational interaction.

Throughout an interaction in the Proving Ground module, students interacted with three agents. These agents include Dr. Quinn (the teacher agent from Cadet Training), Tracy (competing agent) and Broth (a character in the storyline who does not actively interact with other agents and the human). The primary interactions occur between the human student and Dr. Quinn and Tracy. Broth simply provides commentary to help move the story forward.

The human student competes against Tracy to win points in a game requiring research methodology skills. During a typical interaction with the game, the human student chooses a case from a list of potential topics. Next, the student reads the case and must type in a flaw in a case or say that there are no more flaws in the case by pressing a "no flaw" button. If the student is unable to identify a flaw or presses the "no flaw" button when there are flaws in the case, then the next turn goes to Tracy, the competing peer agent. The algorithm behind the game ensures that the competing peer agent will not win the game. This is necessary so that the human student can progress to the third module, the Active Duty module, where the student actively generates questions. With this in mind, during turn-taking in the competition, the competing peer agent is unlikely to immediately provide a correct answer. When Tracy articulates an incorrect answer or misconception, then Dr. Quinn provides a correction. For example, Tracy may say that there was no control group in a study when in fact there was. Dr. Quinn provides negative feedback, such as "no, you are incorrect", followed by a correction such as "in this study, the group of participants who did not receive the manipulation are in the control group". After Tracy has been corrected, it is the human student's turn to identify a flaw. If the human student is unable to articulate a correct flaw, then Tracy takes another turn and may also incorrectly identify a flaw. At this point, Dr. Quinn provides hints and prompts in order to help the human student articulate the correct answer. At any point, the human student can purchase a list of potential flaws by spending score points as well as access the e-text. The human and the artificial agent take turns until no more flaws are found in a case. Throughout the interaction, students are shown their scores indicating their performance, with a higher score indicating better performance. During a typical interaction, the human player must assess approximately 10 or more cases before progressing to the final module, the Active Duty module.

Trialogues during question generation in the Active Duty module

The Active Duty module requires students to actively generate questions about research abstracts. This module requires the skills of question generation (Graesser, Ozuru and Sullins 2009; Graesser, McNamara and VanLehn, 2005) as well as deep-level reasoning and subtle discriminations. The premise of this module is that a potential invader has been captured and it is the human student's job to interrogate the suspect to discover if the research is flawed, thus making the suspect an invader.

The activities in the module primarily consist of the human student asking questions with the aid of Scott, an "interrogator" agent, and the student receiving answers from an unidentified captured invader. In the beginning of each interaction, the human student is only presented with an abstract of the research case. Next the student must ask questions such as "Was there a control group?" and correctly evaluate the suspect's answer. The question is then repeated by the artificial interrogator agent, to model a good question. If the student's question is recognised by the system as an expected question based on the language processing match score, Scott responds with neutral-positive feedback such as "OK" and then asks the student's question. If the student's question matches one that has been previously answered, then Scott responds with somewhat negative feedback (e.g. "Let me ask a different question"). However, if the system is uncertain whether or not the student is asking a good question (i.e. it does not match an expectation or a covered answer), Scott will provide neutral feedback (e.g. "I'm not sure I understand. Let me ask this question instead"). For example, the student may pose a question to a researcher about an abstract that suggests that music helps plants grow. Suppose the question is "How was the dependent variable measured?" The answer may be contextually specific to the case and may not only answer the question at hand but it may answer other generic questions such as "How was the independent variable operationally defined?" Throughout the interaction, students are shown their score indicating their performance in the game.

Assessment of learning

As the students interact with the three modules, the game's log files capture over 1 000 variables about those interactions. Multiple investigations of these variables have been conducted to discover relationships between trialogues in ARIES/ARA and learning and emotions (affect). This section reviews some of these findings and then goes on to discuss correlations in ARIES/ARA responses with students' prior knowledge that may be important for assessment.

Multiple studies have shown significant learning gains from interactions with ARIES/ARA (Halpern et al., 2012; Forsyth et al., 2012, 2013). One consistent theme across studies is that word/idea generation itself is important for learning (Forsyth et al., 2012, 2014; Forsyth, 2014). This phenomenon has been seen in the literature for decades with numerous studies in a variety of environments (Slameka and Graf, 1978; Hirshman and Bjork, 1988; Chi et al., 2001; Rosé et al., 2003), so ARIES is no exception. The trialogues serve an important function of encouraging students' generation of information through pumps, hints and prompts.

During the game, students acquire both shallow, didactic knowledge and also deeper reasoning. The Cadet Training module teaches shallow factual information whereas a combination of the Cadet Training and Proving Ground modules aid in deep-level learning (Forsyth, Graesser, Walker et al., 2013).

These shallow versus deep constructs are important in assessment of learning. For example, while simple word generation is important for learning factual information, discriminating between types of flaws is important for deep-level learning, but prior knowledge is an important mediating variable in this claim (Forsyth et al., 2012, 2014; Forsyth, 2014).

We analysed the student contributions with a computational linguistic tool known as Coh-Metrix (Graesser, McNamara et al., 2014). The tool analyses text on over 100 indices and reduces them to 5 principal components and an overall formality score. The five principle component scores are: word concreteness, syntactic complexity, referential cohesion, deep cohesion and genre. The formality score increases with abstract words, complex syntax, cohesion and the informational genre (as opposed to narrative). Analyses of the inputs students generated during interaction with the game revealed that syntactic complexity and formality of language were both positively correlated with prior knowledge. The more complex and formal the students' discourse was, then the higher their levels of prior knowledge. All together, these findings suggest that student answers containing high-quality contributions that are relevant to the subject matter and have high levels syntactic complexity and formality may be features that are diagnostic of student knowledge levels.

Closing comments

Substantial progress in computer science, artificial intelligence, automated natural language understanding, computational linguistics, discourse processes, intelligent tutoring systems, informatics, text-to-speech generation and avatar design make it possible to create formative assessment environments using natural language conversations. We have seen advances in systems where students converse with one, two or even three artificial agents to collaboratively reason and solve problems. Such systems provide the potential to teach and assess skills that before could only be done using multiple students and time-consuming scoring and monitoring by instructors. This permits students to interact in more realistic collaborative and tutoring situations while receiving immediate, focused feedback.

We are excited to see the continued work of trialogues in assessments being developed at Educational Testing Service where students interact and converse with agents during formative assessment of English language skills, mathematics, science and argumentation. These conversations may make it possible to disentangle multiple constructs such as English language learning skills from mathematics skills. We also look forward to watching the impact both locally with the Center for the Study of Adult Literacy and globally with PISA 2015. Building on prior research including AutoTutor and OperationARIES!, we see how new ventures create new possibilities for formative assessment via conversations with animated pedagogical agents.

References

Anderson, J.R., A.T. Corbett, K.R. Koedinger and R. Pelletier (1995), “Cognitive tutors: Lessons learned”, *Journal of the Learning Sciences*, Vol. 4/2, pp. 167-207.

Azevedo, R., A. Johnson, A. Chauncey and C. Burkett (2010), “Self-regulated learning with MetaTutor: Advancing the science of learning with metacognitive tools”, in M.S. Khine and I.M. Saleh (eds.), *New Science of Learning: Cognition, Computers and Collaboration in Education*, Springer, Amsterdam, pp. 225-247.

Biswas, G., et al. (2010), “Measuring self-regulated learning skills through social interactions in a teachable agent environment”, *Research and Practice in Technology-Enhanced Learning*, Vol. 5/2, pp. 123-152.

Cai, Z., et al. (2014), “Instructional strategies in trialog-based intelligent tutoring systems”, in R. Sottilare, A.C. Graesser, X. Hu and B. Goldberg (eds.), *Design Recommendations for Intelligent Tutoring Systems – Volume 2: Instructional Management*, Army Research Laboratory, Orlando, FL, pp. 225-235.

Cai, Z., et al. (2011), “Trialog in ARIES: User input assessment in an intelligent tutoring system”, in W. Chen and S. Li (eds.), *Proceedings of the 3rd IEEE International Conference on Intelligent Computing and Intelligent Systems,* IEEE Press, Guangzhou, pp. 429-433.

Cai, Z., A.C. Graesser and X. Hu (2015), “ASAT: AutoTutor script authoring tool”, in. R. Sottilare, et al. (eds.), *Design Recommendations for Intelligent Tutoring Systems – Volume 3: Authoring Tools and Expert Modeling Techniques* Army Research Laboratory, Orlando, FL, pp. 199-210.

Chi, M.T.H., M. Roy and R.G. Hausmann (2008), “Observing tutoring collaboratively: Insights about tutoring effectiveness from vicarious learning”, *Cognitive Science*, Vol.32/2, pp. 301-341.

Chi, M.T.H., et al. (2001), “Learning from human tutoring”, *Cognitive Science*, Vol. 25/4, pp. 471-533.

Craig, S.D., B. Gholson, J.K. Brittingham, J.L. Williams and K.T. Shubeck (2012), “Promoting vicarious learning of physics using deep questions with explanations”, *Computers & Education*, Vol. 58/4, pp. 1042-1048.

D’Mello, S.K. and A.C. Graesser (2012), “AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back”, *ACM Transactions on Interactive Intelligent Systems*, Vol. 2/4, pp. 1-39.

Dillenbourg, P. and D. Traum (2006), “Sharing solutions: Persistence and grounding in multi-modal collaborative problem solving”, *Journal of the Learning Sciences*, Vol. 15/1, pp. 121-151.

Driscoll, D.M., S.D. Craig, B. Gholson, M. Ventura, X. Hu and A.C. Graesser (2003), “Vicarious learning: Effects of overhearing dialog and monologue-like discourse in a virtual tutoring session, *Journal of Educational Computing Research*, Vol. 29/4, pp. 431-450.

Dzikovska, M.O., N. Steinhauser, E. Farrow, J.D. Moore and G.E. Campbell (2014), “BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics”, *International Journal of Artificial Intelligence in Education*, Vol. 24/3, pp. 284-332.

Fiore, S. M. and J.W. Schooler (2004), “Process mapping and shared cognition: Teamwork and the development of shared problem models”, in E. Salas and S.M. Fiore (eds.), *Team Cognition: Understanding the Factors that Drive Process and Performance*, American Psychological Association, Washington, DC, pp. 133-152.

Foltz, P.W., W. Kintsch and T.K. Landauer (1998), "The measurement of textual coherence with latent semantic analysis", *Discourse Processes*, Vol. 25/2-3, pp. 285-307.

Foltz, P.W., D. Laham and M. Derr (2003), "Automated speech recognition for modeling team performance", in *Proceedings of the 47th Annual Human Factors and Ergonomic Society Meeting*, Human Factors and Ergonomics Society.

Foltz, P.W. and M.J. Martin (2008), "Automated communication analysis of teams", in E. Salas, G.F. Goodwin and C.S. Burke (eds.), *Team Effectiveness in Complex Organizations: Cross-Disciplinary Perspectives and Approaches*, Routledge, New York.

Forsyth, C.M. (2014), "Predicting learning: A fine-grained analysis of learning in a serious game", unpublished dissertation, University of Memphis.

Forsyth, C.M., et al. (2013), "Operation ARIES!: Methods, mystery and mixed models: Discourse features predict affect in a serious game", *Journal of Educational Data Mining,* Vol. 5/1.

Forsyth, C.M., et al. (2014), "Discovering theoretically grounded predictors of shallow vs. deep- level learning", in J. Stamper, Z. Pardos, M. Mavrikis and B.M. McLaren (eds.), *Proceedings of the 7th International Conference on Educational Data Mining,* Educational Data Mining 2014, pp. 229-232.

Forsyth, C.M., et al. (2013), "Didactic galactic: Types of knowledge learned in a serious game", in H.C. Lane, K. Yacef, J. Mostow and P. Pavlik (eds.), *Artificial Intelligence in Education: 16th International Conference* (AIED 2013), Springer, Heidelberg, pp. 832-835

Forsyth, C.M., et al. (2013), "Assessing performance metrics within a serious game", presentation at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Forsyth, C.M. et al. (2012), "Learning gains for core concepts in a serious game on scientific reasoning", in K. Yacef, et al. (eds.), *Proceedings of the 5th International Conference on Educational Data Mining,* International Educational Data Mining Society, Chania, Greece, pp. 172-175.

Funke, J. (2010), "Complex problem solving: A case for complex cognition?", *Cognitive Processing*, Vol. 11/2, pp. 133-142, http://dx.doi.org/10.1007/s10339-009-0345-0.

Gholson, B. et al. (2009), "Exploring the deep-level reasoning questions effect during vicarious learning among eighth to eleventh graders in the domains ofcomputer literacy and Newtonian physics", *Instructional Science*, Vol. 37/5, pp. 487-493.

Gorman, J.C. et al. (2003), "Evaluation of latent semantic analysis-based measures of team communications content", in *Proceedings of the 47th Annual Human Factors and Ergonomic Society Meeting*, Human Factors and Ergonomics Society.

Graesser, A.C. (2016), "Conversations with AutoTutor help students learn", *International Journal of Artificial Intelligence in Education*, vol-361 pp. 124-132.

Graesser, A.C., et al. (2015). Emotions in adaptive computer technologies for adults improving learning. In S. Tettegah and M. Gartmeier (Eds.), *Emotions, technology, learning, and design,* pp. 2-25. San Diego: Elsevier.

Graesser, A.C., et al. (2016). Reading comprehension lessons in AutoTutor for the Center for the Study of Adult Literacy. In S.A. Crossley and D.S. McNamara (Eds.). *Adaptive educational technologies for literacy instruction* (pp. 288-293) New York: Taylor & Francis Routledge.

Graesser, A. C., S.K. D'Mello and W. Cade (2011), "Instruction based on tutoring", in R .E. Mayer and P.A. Alexander (eds.), *Handbook of Research on Learning and Instruction,* Routledge Press, New York, pp. 408-426.

Graesser, A.C., et al. (2012), "AutoTutor", in P. McCarthy and C. Boonthum-Denecke (eds.), *Applied Natural Language Processing: Identification, Investigation and Resolution,* IGIGlobal, Hershey, PA, pp. 169-187.

Graesser, A.C., et al. (forthcoming), "Challenges of assessing collaborative problem solving", in E. Care, P. Griffin and M. Wilson (eds.), Assessment and Teaching of 21st Century Skills, Springer, Heidelberg.

Graesser, A.C., C. Forsyth and B. Lehman (in press), "Two heads are better than one: Learning from agents in conversational trialogues", *Teacher College Record*.

Graesser, A.C., M. Jeon and D. Dufty (2008), "Agent technologies designed to facilitate interactive knowledge construction", *Discourse Processes*, Vol. 45/4-5, pp. 298-322.

Graesser, A.C., H. Li and C. Forsyth (2014), "Learning by communicating in natural language with conversational agents", *Current Directions in Psychological Science*, Vol. 23/5, pp. 374-380.

Graesser, A.C., et al. (2004), "AutoTutor: A tutor with dialogue in natural language", *Behavioral Research Methods, Instruments, and Computers*, Vol. 36/2, pp. 180-193.

Graesser, A.C., et al. (2014), "Coh-Metrix measures text characteristics at multiple levels of language and discourse", *Elementary School Journal*, Vol. 115/2, pp. 210-229.

Graesser, A.C., D. McNamara and K. VanLehn (2005), "Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iSTART", *Educational Psychologist*, Vol. 40/4, pp. 225-234.

Graesser, A.C., Y. Ozuru and J. Sullins (2009), "What is a good question?", in M.G. McKeown and L. Kucan (eds.), *Threads of Coherence in Research on the Development of Reading Ability*, Guilford Press, New York, pp. 112-141.

Graesser, A.C. and N.K. Person (1994), "Question asking during tutoring", *American Educational Research Journal*, Vol. 31/1, pp. 104-137.

Graesser, A.C., N.K. Person and J.P. Magliano (1995), "Collaborative dialogue patterns in naturalistic one-to-one tutoring", *Applied Cognitive Psychology*, Vol. 9/6, pp. 495-522.

Halpern, D.F. et al. (2012), "Operation ARA: A computerized learning game that teaches critical thinking and scientific reasoning", *Thinking Skills and Creativity*, Vol. 7/2, pp. 93-100.

Hirshman, E. and R.A. Bjork (1988), "The generation effect: Support for a two-factor theory", *Journal of Experimental Psychology: Learning, Memory and Cognition*, Vol. 14/3, pp. 484-494.

Jackson, G.T. and A.C. Graesser (2006), "Applications of human tutorial dialog in AutoTutor: An intelligent tutoring system", *Revista Signos*, Vol. 39/60, pp. 31-48.

Jackson, G.T. and D.S. McNamara (2013), "Motivation and performance in a game-based intelligent tutoring system", *Journal of Educational Psychology*, Vol. 105/4, pp. 1036-1049.

Johnson, W.L. and A. Valente (2008), "Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures", in M. Goker and K. Haigh (eds.), *Proceedings of the Twentieth Innovative Applications of Artificial Intelligence Conference*, AAAI Press, Menlo Park, CA, pp. 1632-1639.

Jurafsky, D. and J. Martin (2008), *Speech and Language Processing*, 2nd edition, Prentice Hall, Englewood, NJ.

Kopp, K.J., et al. (2012), "Improving the efficiency of dialogue in tutoring", *Learning and Instruction*, Vol. 22/5, pp. 320-330.

Lane, H.C. et al. (2011), “Intelligent tutoring goes to the museum in the big city: A pedagogical agent for informal science education”, in G. Biswas, et al. (eds.), *Artificial Intelligence in Education: 15th International Conference* (AEID 2011) Spring, Heidelberg, pp. 155-162.

Landauer, T.K, P.W. Foltz and D. Laham (1998), “An introduction to latent semantic analysis”, *Discourse Processes*, Vol. 25, pp. 259-284.

Landauer, T.K., et al. (2007), *Handbook of Latent Semantic Analysis*, Erlbaum, Mahwah, NJ.

Leelawong, K. and G. Biswas (2008), “Designing learning by teaching agents: The Betty’s Brain system”, *International Journal of Artificial Intelligence in Education*, Vol. 18/3, pp. 181-208.

Lehman, B. et al. (2013), “Inducing and tracking confusion with contradictions during complex learning”, *International Journal of Artificial Intelligence in Education*, Vol. 22/1-2, pp. 85-105.

Mayer, R.E. (2009), *Multimedia Learning,* 2nd edition, Cambridge University Press, New York.

Martin, M.J. and P.W. Foltz (2004), “Automated team discourse annotation and performance prediction using LSA”, in *Proceedings of the Human Language Technology Conference of the NAACL,* HLT/NAACL.

McCarthy, P.M and C. Boonthum-Denecke (eds.) (2012), *Applied Natural Language Processing: Identification, Investigation and Resolution,* IGIGlobal, Hershey, PA.

McNamara, D.S., T. O’Reilly, R. Best and Y. Ozuru (2006), “Improving adolescent students’ reading comprehension with iSTART”, *Journal of Educational Computing Research*, Vol. 34, pp. 147-171.

Millis, K., et al. (2011), “Operation ARIES! A serious game for teaching scientific inquiry”, in M. Ma, A. Oikonomou and J. Lakhmi (eds.), *Serious Games and Edutainment Applications*, Springer-Verlag, London, pp. 169-195.

Nye, B.D., A.C. Graesser and X. Hu (2014), “AutoTutor and family: A review of 17 years of natural language tutoring”, *International Journal of Artificial Intelligence in Education*,Vol. 24/4, pp. 427-469.

OECD (2013), *PISA 2015: Draft Collaborative Problem Solving Framework*, OECD, Paris, www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf.

OECD (2010), *PISA 2012 Field Trial Problem Solving Framework*, OECD, Paris.

Olney, A. et al. (2012), “Guru: A computer tutor that models expert human tutors”, in S. Cerri, W. Clancey, G. Papadourakis and K. Panourgia (eds.), *Intelligent Tutoring Systems: 11th International Conference, ITS 2012, Proceedings,* Springer, Heidelberg, pp. 256-261

Person, N.K. and A.C Graesser (1999), “Evolution of discourse in cross-age tutoring”, in A.M. O’Donnell and A. King (eds.), *Cognitive Perspectives on Peer Learning,* Lawrence Erlbaum, Mahwah, NJ, pp. 69-86.

Person, N.K., et al. (2001), “Simulating human tutor dialog moves in AutoTutor”, *International Journal of Artificial Intelligence in Education*, Vol. 12, pp. 23-39.

PIAAC Expert Group in Problem Solving in Technology-Rich Environments (2009), “PIAAC problem solving in technology-rich environments: A conceptual framework”, *OECD Education Working Papers*, No. 36, OECD Publishing, Paris, http://dx.doi.org/10.1787/220262483674.

Roediger, H.L., M. McDaniel and K. McDermott (2006), “Test enhanced learning”, *APS Observer*, Vol. 19, 28.

Roscoe, R.D., J. Wagster and G. Biswas (2008), “Using teachable agent feedback to support effective learning-by-teaching”, in *Proceedings of the 30th Annual Meeting of the Cognitive Science Society,* Washington, DC, pp. 2381-2386.

Rosé, C.P., D. Bhembe, S. Siler, R. Srivastava and K. VanLehn (2003), "Exploring the effectiveness of knowledge construction dialogues", in U. Hoppe, F. Verdejo and J. Kay (eds.), *Artificial Intelligence in Education: Shaping the Future of Learning through Intelligent Technologies*, IOS Press, Amsterdam, pp. 497-499.

Rosen, Y. and P.W. Foltz (2014), "Assessing collaborative problem solving through automated technologies", *Research and Practice in Technology Enhanced Learning*, Vol. 9/3, pp. 389-410.

Rowe, J., et al. (2010), "Integrating learning, problem solving, and engagement in narrative-centered learning environments", *International Journal of Artificial Intelligence in Education*, Vol. 21, pp. 115-133.

Rus, V., S. D'Mello, X. Hu and A.C. Graesser (2013), "Recent advances in intelligent systems with conversational dialogue", *AI Magazine*, Vol. 34/3, pp. 42-54.

Slamecka, N.J. and P. Graf (1978), "The generation effect: Delineation of a phenomenon", *Journal of Experimental Psychology: Human Learning and Memory*, Vol. 4/6, pp. 592-604.

VanLehn, K. (2011), "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems", *Educational Psychologist*, Vol. 46/4, pp. 197-221.

VanLehn, K. (2006), "The behavior of tutoring systems", *International Journal of Artificial Intelligence in Education*, Vol. 16/3, pp. 227-265.

VanLehn, K. et al. (2007), "When are tutorial dialogues more effective than reading?", *Cognitive Science*, Vol. 31, pp. 3-62.

Ward, W. et al. (2013), "My science tutor: A conversational multimedia virtual tutor", *Journal of Educational Psychology*, Vol. 105/4, pp. 1115-1125.

Zapata-Rivera, D., G.T. Jackson and I. Katz (2015), "Authoring conversation-based assessment scenarios", in. R. Sottilare, A.C. Graesser, X. Hu and K. Brawner (eds.), *Design Recommendations for Intelligent Tutoring Systems – Volume 3: Authoring Tools and Expert Modeling Techniques* Army Research Laboratory, Orlando, FL, pp. 169-178.

PART VI

Finale

Chapter 16

Epilogue

By
Benő Csapó
MTA-SZTE Research Group on the Development of Competencies, University of Szeged, Hungary.
and Joachim Funke
Department of Psychology, Heidelberg University, Germany.

As we finish the work of editing this book and look back at the process behind that effort, we have the feeling that it has simultaneously been one of the most inspiring experiences and one of the most challenging undertakings of our professional careers. The idea of creating a book on the state of the art of problem solving research was conceived in a series of formal and informal meetings among the authors, and it has been shaped in a number of discussions and e-mail exchanges since then.

Editors of collected studies always begin their endeavor with clear goals, and when they read the first drafts of the chapters, they often suggest revisions to reach those goals. At the same time, they may also revise their original intentions. During the editing process, we did so in two respects. On the one hand, the materials we have received represent a much broader perspective on the field and cover a number of recent aspects of problem-solving research that were not visible at the beginning of the project.

On the other hand, at the beginning we envisioned a well-structured, textbook-like volume with clear definitions of concepts and a consistent use of terminology throughout the book. However, over the past years, we have had to revise our own original ideas, as it has become obvious that the use of the terminology is so diverse that the way the authors use it is far from consistent.

When we attempted to establish a uniform terminology, we realised that proposing the terms we prefer would limit the authors in expressing their ideas; furthermore, we knew they would use their preferred terminology in their other publications anyway. Such a strong editorial approach might have made this book terminologically more consistent, but it would have created inconsistency between the authors' present and other publications and would not have aided in making the terms used in the field any clearer. Therefore, we accepted the authors' terminology as they use it to express their ideas in their own way, but in this concluding chapter, we discuss this issue and propose a conceptual system only for the most crucial issues.

The problem of the terminology for problem solving is not unique at all. The difficulty becomes clear if we think of the disputes surrounding interpretations of some other general terms, such as knowledge, skills, competencies, intelligence, creativity, aptitude and ability, and the relationship between them. There are a number of factors that make it difficult to develop a consistent terminology. One is the historical issue. As we discussed in the first chapters, the conception of problem solving has evolved over a century, and different schools and traditions have used different sets of terms that are not always easy to map onto each other. In recent decades, problem-solving research has expanded at an accelerating speed. At this stage of rapid development when researchers identify and define its ever newer forms, it would not be practical (if even possible) to freeze the current state of categorisation or propose a rigid terminology.

Another issue is a linguistic one, as problem or problem solving may have different connotations in different languages. In different languages, the sections of the continuum that spans from carrying out routine tasks to solving complex problems may be referred to differently, and the point from where the activity is called problem solving may be found at different places. For example, problem in terms of "mathematical word problem" would be translated as *Aufgabe* into German or *feladat* into Hungarian, although a cognate for the word *problem* exists in both languages. Some languages have different terms for the simpler forms of problem solving, reserving their version of the term *problem solving* for the more complex forms and thus no adjective is needed to express its complex nature. These different linguistic tools shape our thinking about problem solving in different ways, which also contribute to the difficulty in using a consistent terminology. Nevertheless, we have to notice that although the usage of the terms throughout this book does not always seem consistent, the way of thinking across the authors of the chapters seems very consistent indeed.

The terms describing types of problem solving we have used in the chapters of this book may best be characterised by five pairs of antinomies:

1) analytical/complex
2) static/dynamic
3) individual/collaborative
4) domain-specific/domain-general
5) subject-specific/cross-curricular.

The nature of these antinomies has been elaborated in several chapters and in different contexts, so we think it is unnecessary to go into detail here. But at least a short summary of our understanding of the dichotomies will be presented here.

We propose the "analytical/complex" dichotomy as the most general division of the categories of problem solving. Needless to say, we consider the complex side the more interesting and challenging field of research, and this book tends to deal with that. Analytical problems are closed problems in the sense that all the necessary information that a problem solver needs to arrive at a solution is given. In complex problems, there is no clear-cut border to the problem situation.

The "static/dynamic" dichotomy is highlighted when the nature of dynamic problem solving is explained, and static in this context simply means not dynamic. Dynamic problem solving has become the focal area of research recently, as it can best be realised by means of computer simulation. We consider dynamic problem solving the most advanced form, and thus a subcategory of complex problem solving. Dynamic problem solving always involves an interactive knowledge acquisition phase. Therefore, in certain contexts it may be called interactive problem solving when its interactive nature is emphasised; however, in general, the term dynamic is more precise. There is also a great deal of overlap between the connotations of static and analytical; it is the context that determines which one is more appropriate.

The "individual/collaborative" dichotomy indicates the social context of problem solving and also describes a relatively new field of research. Collaborative problem solving also involves interactions, in this case those between problem-solving individuals. Collaborative problem solving originally denoted human-to-human interactions, although there have been attempts to replace one of the humans with agents (computer-simulated collaborators). Such simulated environments may be used for assessment purposes, although certain aspects of real social interactions may thus be lost. Some facets of reality are also lost when simulation is used in dynamic problem solving, and so one of the most interesting issues in current research is how well human-agent interaction represents human-human interactions and what the relationship between these and the process of dynamic problem solving is.

The "domain-specific/domain-general" dichotomy indicates the scope of the problem-solving skills, with the domain-general side tending to be closer to the complex side of the "analytical/complex" dichotomy discussed above. The "subject-specific/cross-curricular" division reflects the same issue in an educational context.

Several combinations of these adjectives may be reasonable. For example, we may deal with domain-general dynamic problem solving or domain-specific static problem-solving skills. A number of other adjectives may also be attached to problem solving to highlight some of its specific contexts or aspects, for example scientific problem solving. The volume that reported problem solving for PISA 2012 was entitled "Creative problem solving", thus highlighting the creative nature of dynamic problem solving, which had been included in the assessment instruments.

We have witnessed the birth of a new level of problem-solving research. We have also seen ever newer fields being made measurable in the domain of problem solving in a broader sense; indeed, one of the most interesting goals of research is to make measurable that which cannot yet be measured. Educational assessment provides vital feedback for several levels of education systems. PISA 2003 introduced the assessment of problem solving to large-scale international programs, and the 2012 assessment brought to light its broader potential in evaluating the quality of students' knowledge and skills. We hope with time that PISA will consider incorporating problem solving into its assessments in innovative domains when country differences may change due to their efforts to improve the quality of education. We also believe that as the feasibility of assessment is demonstrated at an international level, it will inspire countries to include an assessment of problem solving in their national assessment systems, as the testing of reading, mathematical and scientific literacy did. Technology-based assessment of problem solving may also pave the way for developing student-level feedback systems as well.

Precise and reliable assessment is a precondition for the empirical testing of a number of exciting hypotheses. There is a growing consensus among researchers of cognition that the most important general abilities are plastic; their development may be accelerated and extended if the developing mind is stimulated with proper exercises at the right time. By reviewing current research results, we may conclude that problem solving is amenable, but so far there has been too little research on how it can be improved in school contexts. Making this domain measurable with easy-to-use instruments may give new impetus to well-controlled interventions, thus strengthening the scientific foundations of development in education. We envision that training experiments will create the most promising domain of future research on problem solving.

Contributors

Ingo Barkow, Swiss Institute for Information Science (SII), University of Applied Sciences HTW Chur, Pulvermühlestrasse 57, 7004 Chur, Switzerland; ingo.barkow@htwchur.ch

Martin Brunner, FU Berlin, Institut für Schulqualität, Otto-von-Simson-Str. 15, 14195 Berlin, Germany; sekbrunner@zedat.fu-berlin.de

Florian Buchwald, Faculty of Physics, Physics Education, University of Duisburg-Essen, 45117 Essen, Germany; florian.buchwald@uni-due.de

Esther Care, Brookings Institution, 1775 Massachusetts Ave NW, Washington, DC 20036, United States; ecare@brookings.edu

Benő Csapó, MTA-SZTE Research Group on the Development of Competencies,University of Szeged, Petőfi sgt. 30–34, 6722 Szeged, Hungary; csapo@edpsy.u-szeged.hu

John A. Dossey, Distinguished Professor of Mathematics Emeritus, Illinois State University, 65111 East Crystal Ridge, Saddle Brooke, AZ 85739, United States; jdossey44@gmail.com

Andreas Fischer, Forschungsinstitut Betriebliche Bildung (f-bb), Rollnerstraße 14, 90408 Nürnberg, Germany; andreas.fischer@f-bb.de

Jens Fleischer, Faculty of Educational Science, Department of Instructional Psychology, University of Duisburg-Essen, 45117 Essen, Germany; jens.fleischer@uni-due.de

Peter Foltz, University of Colorado at Boulder and Pearson, 4940 Pearl East Circle, Boulder, CO 80301, United States; pfoltz@pearsonkt.com

Carol M. Forsyth, Educational Testing Service, 90 New Montgomery Street, Suite 1500, San Francisco, CA 94105, United States; cforsyth@ets.org

Annemarie Fritz, University of Essen, Institut für Psychologie, Fakultät für Bildungswissenschaften, Berliner Platz 6-8, 45117 Essen, Germany; fritz-stratmann@uni-due.de

Joachim Funke, Psychologisches Institut, Universität Heidelberg, Hauptstr. 47, 69117 Heidelberg, Germany; joachim.funke@psychologie.uni-heidelberg.de

Frank Goldhammer, German Institute for International Educational Research, Schloßstraße 29, 60486 Frankfurt am Main, Germany; goldhammer@dipf.de

Arthur C. Graesser, Psychology Department, 202 Psychology Building, University of Memphis, Memphis, TN 38152–3230, US; graesser@memphis.edu

Samuel Greiff, Université du Luxembourg, Maison des Sciences Humaines, 11 Porte des Sciences, 4366 Esch-sur-Alzette, Luxembourg; samuel.greiff@uni.lu

Patrick Griffin, Assessment Research Centre, University of Melbourne, Victoria 3010, Australia; p.griffin@unimelb.edu.au

Ulrich Keller, Luxembourg Centre of Educational Testing, University of Luxembourg, Maison des Sciences Humaines, 11 Porte des Sciences, 4366 Esch-sur-Alzette, Luxembourg; ulrich.keller@uni.lu

Thibaud Latour, Luxembourg Institute of Science and Technology, Luxembourg; thibaud.latour@list.lu

Detlev Leutner, Faculty of Educational Science, Department of Instructional Psychology, University of Duisburg-Essen, 45117 Essen, Germany; detlev.leutner@uni-due.de

Romain Martin, Luxembourg Centre of Educational Testing, University of Luxembourg, Maison des Sciences Humaines, 11 Porte des Sciences, 4366 Esch-sur-Alzette, Luxembourg; romain.martin@uni.lu

Barry Mccrae, The University of Melbourne and Australian Council for Educational Research, 19 Prospect Hill Road, Camberwell VIC 3124, Australia; barry.mccrae@acer.edu.au

Gyöngyvér Molnár, MTA-SZTE Research Group on the Development of Competencies, Petöfi sgt. 30–34, 6722 Szeged, Hungary; gymolnar@edpsy.u-szeged.hu

Magda Osman, Experimental Cognitive Psychology, Department of Psychology, School of Biological and Chemical Sciences, Queen Mary University of London, Mile End, London E1 4NS, United Kingdom; m.osman@qmul.ac.uk

Ray Philpot, Australian Council for Educational Research, 19 Prospect Hill Road, Camberwell VIC 3124, Australia; ray.philpot@acer.edu.au

Dara Ramalingam, Australian Council for Educational Research, 19 Prospect Hill Road, Camberwell VIC 3124, Australia; dara.ramalingam@acer.edu.au

Heiko Rölke, German Institute for International Educational Research, Schloßstraße 29, 60486 Frankfurt am Main, Germany; roelke@dipf.de

Stefan Rumann, Faculty of Chemistry, Institute of Chemistry Education, University of Duisburg-Essen, Schützenbahn 70, 45127 Essen, Germany; stefan.rumann@uni-due.de

Andreas Schleicher, OECD, 2 rue André Pascal, 75775 Paris Cedex 16, France; andreas.schleicher@oecd.org

Philipp Sonnleitner, Luxembourg Centre of Educational Testing, University of Luxembourg, Maison des Sciences Humaines, 11 Porte des Sciences, 4366 Esch-sur-Alzette, Luxembourg; philipp.sonnleitner@uni.lu

David Tobinski, University of Essen, Institut für Psychologie, Fakultät für Bildungswissenschaften, Berliner Platz 6–8, 45117 Essen, Germany; david.tobinski@uni-due.de

Krisztina Tóth, Szeged University, Center for Research on Learning and Instruction, Petöfi sgt. 30–34, 6722 Szeged, Hungary: and German Institute for International Educational Research, Frankfurt am Main, Germany, Schloßstraße 29, 60486.; ktoth@tba21.hu

Joachim Wirth, Institute of Educational Sciences, Department of Research on Learning and Instruction, Ruhr-University Bochum, Universitätsstraße 150, 44780 Bochum, Germany; lehrlernforschung@rub.de

Sascha Wüstenberg, TWT GmbH Science & Innovation, Ernsthaldenstrasse 17, 70565 Stuttgart, Germany; sascha.wuestenberg@t-online.de

Nathan Zoanetti, Victorian Curriculum and Assessment Authority, Level 1, 2 Lonsdale Street, Melbourne, Victoria, 3000, Australia; zoanetti.nathan.p@edumail.vic.gov.au

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

The OECD is a unique forum where governments work together to address the economic, social and environmental challenges of globalisation. The OECD is also at the forefront of efforts to understand and to help governments respond to new developments and concerns, such as corporate governance, the information economy and the challenges of an ageing population. The Organisation provides a setting where governments can compare policy experiences, seek answers to common problems, identify good practice and work to co-ordinate domestic and international policies.

The OECD member countries are: Australia, Austria, Belgium, Canada, Chile, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Latvia, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, the United Kingdom and the United States. The European Union takes part in the work of the OECD.

OECD Publishing disseminates widely the results of the Organisation's statistics gathering and research on economic, social and environmental issues, as well as the conventions, guidelines and standards agreed by its members.

OECD PUBLISHING, 2, rue André-Pascal, 75775 PARIS CEDEX 16
(96 2017 02 1 P) ISBN 978-92-64-27389-4 - 2017

Educational Research and Innovation

The Nature of Problem Solving

USING RESEARCH TO INSPIRE 21ST CENTURY LEARNING

Solving non-routine problems is a key competence in a world full of changes, uncertainty and surprise where we strive to achieve so many ambitious goals. But the world is also full of solutions because of the extraordinary competences of humans who search for and find them. We must explore the world around us in a thoughtful way, acquire knowledge about unknown situations efficiently, and apply new and existing knowledge creatively.

The Nature of Problem Solving presents the background and the main ideas behind the development of the PISA 2012 assessment of problem solving, as well as results from research collaborations that originated within the group of experts who guided the development of this assessment. It illustrates the past, present and future of problem-solving research and how this research is helping educators prepare students to navigate an increasingly uncertain, volatile and ambiguous world.

Consult this publication on line at: ***http://dx.doi.org/10.1787/9789264273955-en***

This work is published on the *OECD iLibrary,* which gathers all OECD books, periodicals and statistical databases.
Visit ***www.oecd-ilibrary.org*** for more information.

OECD*publishing*
www.oecd.org/publishing

ISBN 978-92-64-24753-6
96 2017 02 1 P

9 789264 247536