

# COMPLEX PROBLEM SOLVING BEYOND THE PSYCHOMETRIC APPROACH

EDITED BY: Wolfgang Schoppek, Joachim Funke, Magda Osman and  
Annette Kluge

PUBLISHED IN: Frontiers in Psychology



# frontiers

## Frontiers Copyright Statement

© Copyright 2007-2018 Frontiers Media SA. All rights reserved.

All content included on this site, such as text, graphics, logos, button icons, images, video/audio clips, downloads, data compilations and software, is the property of or is licensed to Frontiers Media SA ("Frontiers") or its licensees and/or subcontractors. The copyright in the text of individual articles is the property of their respective authors, subject to a license granted to Frontiers.

The compilation of articles constituting this e-book, wherever published, as well as the compilation of all other content on this site, is the exclusive property of Frontiers. For the conditions for downloading and copying of e-books from Frontiers' website, please see the Terms for Website Use. If purchasing Frontiers e-books from other websites or sources, the conditions of the website concerned apply.

Images and graphics not forming part of user-contributed materials may not be downloaded or copied without permission.

Individual articles may be downloaded and reproduced in accordance with the principles of the CC-BY licence subject to any copyright or other notices. They may not be re-sold as an e-book.

As author or other contributor you grant a CC-BY licence to others to reproduce your articles, including any graphics and third-party materials supplied by you, in accordance with the Conditions for Website Use and subject to any copyright notices which you include in connection with your articles and materials.

All copyright, and all rights therein, are protected by national and international copyright laws.

The above represents a summary only. For the full conditions see the Conditions for Authors and the Conditions for Website Use.

ISSN 1664-8714

ISBN 978-2-88945-573-7

DOI 10.3389/978-2-88945-573-7

## About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: [researchtopics@frontiersin.org](mailto:researchtopics@frontiersin.org)



# COMPLEX PROBLEM SOLVING BEYOND THE PSYCHOMETRIC APPROACH

Topic Editors:

**Wolfgang Schoppek**, Universitat Bayreuth, Germany

**Joachim Funke**, Universität Heidelberg, Germany

**Magda Osman**, Queen Mary University of London, United Kingdom

**Annette Kluge**, Ruhr-Universität Bochum, Germany



"Connectivity" by Clint Adair. License: <https://unsplash.com/@clintadair>

Complex problem solving (CPS) and related topics such as dynamic decision-making (DDM) and complex dynamic control (CDC) represent multifaceted psychological phenomena. In a broad sense, CPS encompasses learning, decision-making, and acting in complex and dynamic situations. Moreover, solutions to problems that people face in such situations are often generated in teams or groups. This adds another layer of complexity to the situation itself because of the emerging issues that arise from the social dynamics of group interactions. This framing of CPS means that it is not a single construct that can be measured by using a particular type of CPS task (e.g. minimal complex system tests), which is a view taken by the psychometric community. The proposed approach taken here is that because CPS is multifaceted, multiple approaches need to be taken to fully capture and understand what it is and how the different cognitive processes associated with it complement each other.

Thus, this Research Topic is aimed at showcasing the latest work in the fields of CPS, as well as DDM and CDC that takes a holist approach to investigating and theorizing about these abilities. The collection of articles encompasses conceptual approaches as well as experimental and correlational studies involving established or new tools to examine CPS, DDM and CDC. This work contributes to answering questions about what strategies and what general knowledge can be transferred from one type of complex and dynamic situation to another, what learning conditions result in transferable knowledge and skills, and how these features can be trained.

**Citation:** Schoppek, W., Funke, J., Osman, M., Kluge, A., eds. (2018). Complex Problem Solving Beyond the Psychometric Approach. Lausanne: Frontiers Media.  
doi: 10.3389/978-2-88945-573-7



# Table of Contents

- 05 Editorial: Complex Problem Solving Beyond the Psychometric Approach**  
Wolfgang Schoppek, Annette Kluge, Magda Osman and Joachim Funke

## CONCEPTIONS

- 08 Complex Problem Solving: What it is and What it is Not**  
Dietrich Dörner and Joachim Funke
- 19 The Role of Motivation in Complex Problem Solving**  
C. Dominik Güss, Madison Lee Burger and Dietrich Dörner
- 24 Approaches to Cognitive Modeling in Dynamic Systems Control**  
Daniel V. Holt and Magda Osman
- 30 Evaluation-Dependent Representation in Risk Defusing**  
Oswald Huber

## CORRELATIONAL RESEARCH

- 34 Impact of Cognitive Abilities and Prior Knowledge on Complex Problem Solving Performance – Empirical Results and a Plea for Ecologically Valid Microworlds**  
Heinz-Martin Süß and André Kretzschmar
- 56 The Efficacy and Development of Students' Problem-Solving Strategies During Compulsory Schooling: Logfile Analyses**  
Gyöngyvér Molnár and Benő Csapó
- 73 Potential for Assessing Dynamic Problem-Solving at the Beginning of Higher Education Studies**  
Benő Csapó and Gyöngyvér Molnár
- 85 The Association Between Motivation, Affect, and Self-regulated Learning When Solving Problems**  
Martine Baars, Lisette Wijnia and Fred Paas
- 97 Complex Problem Solving in Teams: The Impact of Collective Orientation on Team Process Demands**  
Vera Hagemann and Annette Kluge

## EXPERIMENTAL RESEARCH

- 114 Beyond Psychometrics: The Difference Between Difficult Problem Solving and Complex Problem Solving**  
Jens F. Beckmann, Damian P. Birney and Natassia Goode
- 127 Common Process Demands of Two Complex Dynamic Control Tasks: Transfer is Mediated by Comprehensive Strategies**  
Wolfgang Schoppek and Andreas Fischer
- 140 Experience in a Climate Microworld: Influence of Surface and Structure Learning, Problem Difficulty, and Decision Aids in Reducing Stock-Flow Misconceptions**  
Medha Kumar and Varun Dutt
- 159 A Cognitive Modeling Approach to Strategy Formation in Dynamic Decision Making**  
Sabine Prezenski, André Brechmann, Susann Wolff and Nele Russwinkel



# Editorial: Complex Problem Solving Beyond the Psychometric Approach

Wolfgang Schoppek<sup>1\*</sup>, Annette Kluge<sup>2</sup>, Magda Osman<sup>3</sup> and Joachim Funke<sup>4</sup>

<sup>1</sup> University of Bayreuth, Bayreuth, Germany, <sup>2</sup> Ruhr-Universität Bochum, Bochum, Germany, <sup>3</sup> School of Biological and Chemical Sciences, Queen Mary University of London, London, United Kingdom, <sup>4</sup> Universität Heidelberg, Heidelberg, Germany

**Keywords:** complex problem solving, dynamic decision making, system control, knowledge acquisition, thinking and reasoning

## Editorial on the Research Topic

### Complex Problem Solving Beyond the Psychometric Approach

In 2005, Quesada et al. (2005) titled a paper “Complex problem-solving: A field in search of a definition.” Thirteen years later, it seems that the field has found it in the form of multiple (or minimal) complex systems such as MicroDYN. While MicroDYN certainly has brought the field a boost of attention and serves as a standard of comparison, not all researchers agree that it is an appropriate operationalization for complex problem solving (CPS). So is the field still searching? A pessimist would affirm this. However, the present collection of articles shows that the search is productive. For the intended scope of the research topic, please refer to the overview.

First, there is a number of conceptual papers. Dörner and Funke stake out the domain of complex problem solving. They claim that complex real-life problems need to be an important part of it. To bridge the gap between internal and ecological validity they recommend utilizing a broad range of research methods. Güss et al. make an argument for incorporating motivation into theoretical considerations about CPS. They substantiate their claim with the analysis of a thinking-aloud protocol of a subject (having worked on the WINFIRE simulation) with respect to three theories of motivation. Holt and Osman give an overview of various approaches to cognitive modeling of dynamic system control. They present strengths and weaknesses of those and conclude that due to the limitations of each single approach hybrid models are most promising. Huber presents theoretical considerations and reviews results about representation and evaluation in decision making. He had identified an advantages first principle, which is cushioned with the use of risk defusing operators. While this research has been conducted in the context of classical decision-making, it appears worthwhile to incorporate its principles into models of CPS. Overall, these papers give a good impression of the diversity of research questions and approaches within CPS and on its borders. Classifying these four papers as conceptual does not mean that the other papers are devoted to pure empiricism. Many of those contain elaborated forms of theoretical reasoning.

A next class of papers can be characterized by using a correlational methodology. Süß and Kretschmar investigated the significance of intelligence and knowledge for performance in two different microworlds: a complex real-life oriented system (Tailorshop) that is largely in line with common knowledge, and a complex artificial world problem (FSYS) that was designed to minimize the influence of prior knowledge. The authors interpret their results as evidence that there is little reason to assume the existence of an ability construct CPS that explains variance in control performance over and above knowledge and reasoning. Molnár and Csapó present a large cross-sectional data set of strategy use in MicroDYN. They classified knowledge acquisition strategies with respect to effectiveness ex ante and found the expected developmental effects. The finding that using effective strategies, although being a predictor of control

## OPEN ACCESS

### Edited and reviewed by:

Eldad Yechiam,  
Technion–Israel Institute of  
Technology, Israel

### \*Correspondence:

Wolfgang Schoppek  
wolfgang.schoppek@uni-bayreuth.de

### Specialty section:

This article was submitted to  
Cognitive Science,  
a section of the journal  
Frontiers in Psychology

**Received:** 04 June 2018

**Accepted:** 27 June 2018

**Published:** 17 July 2018

### Citation:

Schoppek W, Kluge A, Osman M and  
Funke J (2018) Editorial: Complex  
Problem Solving Beyond the  
Psychometric Approach.  
Front. Psychol. 9:1224.  
doi: 10.3389/fpsyg.2018.01224

performance, is nonetheless neither necessary nor sufficient for high performance, shows that things are more intricate as desired, even in the rather plain environment of MicroDYN. In another large study with first year university students, Csapó and Molnár correlated several academic test scores, self-report measures of learning strategies, and MicroDYN performance. Besides the well-established relations among math scores, science scores, and control performance, the study revealed significant paths from elaborative (+) and memorizing (–) learning strategies to math and science test scores. Baars et al. present a study about the involvement of several affective and motivational variables on self-regulated learning of complex hereditary problems. Surprisingly, they found correlations mainly with negative variables: Negative affect, perceived mental effort and low self-assessment accuracy predicted low posttest problem-solving performance, whereas autonomous motivation was not a significant predictor. These results show that it is advisable to assist novice problem-solvers to regulate their mood when it comes to realistic self-assessment. The contribution of Hagemann and Kluge is the only one that addresses CPS in teams. This is an important aspect of solving problems in the real world. The requirement to coordinate team activities can be a source of additional complexity. In the context of a “model of the idealized teamwork process,” the authors investigated the hypothesized relations of cohesion, trust, and collective orientation and found that collective orientation had an effect on team performance mediated by the teams’ coordination behavior.

In a third set of contributions, authors used experiments as their primary method. Beckmann et al. argue, based on a person task situation (PTS) framework, that it is important to distinguish difficulty from complexity. They demonstrate the implications of their claim with an experiment that varied semantic content of the system to be controlled and the assignment of tasks. The results confirmed the expected effects of complexity (induced by situation and task) on the observed difficulty. Schoppek and Fischer compared MicroDYN with a set of more dynamic, real-time driven control tasks (Dynamis2) in a transfer experiment. Besides the expected correlations among control performances in the two tasks and figural reasoning, the experiment revealed positive transfer from MicroDYN to Dynamis2, which was mediated by the use of a specific variant of the VOTAT strategy. Kumar and Dutt tested the effects of a dynamic climate change simulator (a stock-flow scenario) on misconceptions about CO<sub>2</sub> accumulation in the atmosphere. They showed under various conditions that working with the dynamic climate change simulator reduced the misconceptions “correlation heuristic” and “violation of mass balance.” Prezenski et al. present an ACT-R model of a multidimensional classification task. The model combines an exemplar-based and a rule-based approach. It includes perceptual-motor and metacognitive aspects and is able to reproduce the essential effects of the underlying experiment.

At first sight, the diversity of research questions and methods in these contributions seems to be a severe hindrance to drawing conclusions. However, there are crosslinks that can give orientation in this maze.

1. Four papers agree in their appraisal that *motivational variables and processes* need to be considered in the investigation of CPS (Baars et al.; Dörner and Funke; Güss et al.; Hagemann and Kluge). This claim is as plausible as it is hard to prove: Studies have often failed to demonstrate pronounced effects of (global) motivational variables. Therefore, researchers need to incur the laborious approach of identifying and measuring proximal motivational variables that accompany the problem solving process.
2. Several papers address the role of *knowledge in CPS*. The considered knowledge types range from learning instances (e.g., Kumar and Dutt) to explicit knowledge about the structure of a system (e.g., Csapó and Molnár). The significance of prior knowledge is ambiguous: It can be helpful for successful problem solving (Süß and Kretzschmar), but also detrimental (Beckmann et al.). Although “knowledge” is one of the most elaborate topics within CPS, a model that unites several knowledge types in one framework is still a distant prospect (but see Schoppek and Fischer).
3. The question about *strategies for CPS* is closely related to the previous topic, because many of the strategies being investigated in the present collection of papers are serving the purpose of knowledge acquisition: VOTAT (Molnár and Csapó), VONAT (Beckmann et al.), and PULSE (Schoppek and Fischer). In addition, Huber’s advantages first principle and Kumar and Dutt’s correlation heuristic remind us of the possibility that in CPS not every course of action is so clearly scripted as in VOTAT.
4. As *cognitive models of CPS* need to specify knowledge and strategy, they embrace the previous two topics. Even though cognitive models have been used in that context for a while (e.g., Fum and Stocco, 2003), we believe that this methodology still helps to gain a better understanding of the interplay of strategies and the various knowledge types.
5. A final crosslink is constituted by the use of the minimal complex system MicroDYN (Csapó and Molnár; Molnár and Csapó; Schoppek and Fischer), showing that even though many authors take a skeptical view toward MicroDYN as the operationalization of CPS per se, it is still useful as a well-documented point of reference.

We hope that the present collection of articles will stimulate the exchange among researchers in the field of CPS, which is necessary to overcome potential separation. We also hope that it will serve as a guidepost on the way to an architecture of complex problem solving. Summarizing the present proposals, such a differentiated architecture would be hybrid and hierarchical, in order to incorporate diverse elements such as instance-based learning, rule induction, decision-making, and motivational variables.

## AUTHOR CONTRIBUTIONS

WS wrote the first draft of the editorial. All other authors commented on this draft and contributed improved or additional text.



## REFERENCES

- Fum, D., and Stocco, A. (2003). "Instance vs. rule-based learning in controlling a dynamic system," in *Proceedings of the 5th International Conference on Cognitive Modelling*, eds F. Detje, D. Dörner, and H. Schaub (Bamberg: Universitäts-Verlag), 105–110.
- Quesada, J., Kintsch, W., and Gomez, E. (2005). Complex problem solving: a field in search of a definition? *Theor. Issues Ergon. Sci.* 6, 5–33. doi: 10.1080/14639220512331311553

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Schoppek, Kluge, Osman and Funke. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Complex Problem Solving: What It Is and What It Is Not

Dietrich Dörner<sup>1</sup> and Joachim Funke<sup>2\*</sup>

<sup>1</sup> Department of Psychology, University of Bamberg, Bamberg, Germany, <sup>2</sup> Department of Psychology, Heidelberg University, Heidelberg, Germany

Computer-simulated scenarios have been part of psychological research on problem solving for more than 40 years. The shift in emphasis from simple toy problems to complex, more real-life oriented problems has been accompanied by discussions about the best ways to assess the process of solving complex problems. Psychometric issues such as reliable assessments and addressing correlations with other instruments have been in the foreground of these discussions and have left the content validity of complex problem solving in the background. In this paper, we return the focus to content issues and address the important features that define complex problems.

**Keywords:** complex problem solving, validity, assessment, definition, MicroDYN

## OPEN ACCESS

### Edited by:

Andrea Bender,  
University of Bergen, Norway

### Reviewed by:

Bertolt Meyer,  
Technische Universität Chemnitz,  
Germany  
Rumen I. Iliev,  
University of Michigan, United States

### \*Correspondence:

Joachim Funke  
joachim.funke@psychologie.uni-  
heidelberg.de

### Specialty section:

This article was submitted to  
Cognitive Science,  
a section of the journal  
Frontiers in Psychology

**Received:** 14 March 2017

**Accepted:** 23 June 2017

**Published:** 11 July 2017

### Citation:

Dörner D and Funke J (2017)  
Complex Problem Solving: What It Is  
and What It Is Not.  
Front. Psychol. 8:1153.  
doi: 10.3389/fpsyg.2017.01153

Succeeding in the 21st century requires many competencies, including creativity, life-long learning, and collaboration skills (e.g., National Research Council, 2011; Griffin and Care, 2015), to name only a few. One competence that seems to be of central importance is the ability to solve complex problems (Mainzer, 2009). Mainzer quotes the Nobel prize winner Simon (1957) who wrote as early as 1957:

The capacity of the human mind for formulating and solving complex problems is very small compared with the size of the problem whose solution is required for objectively rational behavior in the real world or even for a reasonable approximation to such objective rationality. (p. 198)

The shift from well-defined to ill-defined problems came about as a result of a disillusion with the “general problem solver” (Newell et al., 1959): The general problem solver was a computer software intended to solve all kind of problems that can be expressed through well-formed formulas. However, it soon became clear that this procedure was in fact a “special problem solver” that could only solve well-defined problems in a closed space. But real-world problems feature open boundaries and have no well-determined solution. In fact, the world is full of wicked problems and clumsy solutions (Verweij and Thompson, 2006). As a result, solving well-defined problems and solving ill-defined problems requires different cognitive processes (Schraw et al., 1995; but see Funke, 2010).

Well-defined problems have a clear set of means for reaching a precisely described goal state. For example: in a match-stick arithmetic problem, a person receives a false arithmetic expression constructed out of matchsticks (e.g.,  $IV = III + III$ ). According to the instructions, moving one of the matchsticks will make the equations true. Here, both the problem (find the appropriate stick to move) and the goal state (true arithmetic expression; solution is:  $VI = III + III$ ) are defined clearly.

Ill-defined problems have no clear problem definition, their goal state is not defined clearly, and the means of moving towards the (diffusely described) goal state are not clear.

For example: The goal state for solving the political conflict in the near-east conflict between Israel and Palestine is not clearly defined (living in peaceful harmony with each other?) and even if the conflict parties would agree on a two-state solution, this goal again leaves many issues unresolved. This type of problem is called a “complex problem” and is of central importance to this paper. All psychological processes that occur within individual persons and deal with the handling of such ill-defined complex problems will be subsumed under the umbrella term “complex problem solving” (CPS).

Systematic research on CPS started in the 1970s with observations of the behavior of participants who were confronted with computer simulated microworlds. For example, in one of those microworlds participants assumed the role of executives who were tasked to manage a company over a certain period of time (see Brehmer and Dörner, 1993, for a discussion of this methodology). Today, CPS is an established concept and has even influenced large-scale assessments such as PISA (“Programme for International Student Assessment”), organized by the Organization for Economic Cooperation and Development (OECD, 2014). According to the World Economic Forum, CPS is one of the most important competencies required in the future (World Economic Forum, 2015). Numerous articles on the subject have been published in recent years, documenting the increasing research activity relating to this field. In the following collection of papers we list only those published in 2010 and later: *theoretical* papers (Blech and Funke, 2010; Funke, 2010; Knauff and Wolf, 2010; Leutner et al., 2012; Selten et al., 2012; Wüstenberg et al., 2012; Greiff et al., 2013b; Fischer and Neubert, 2015; Schoppek and Fischer, 2015), papers about *measurement issues* (Danner et al., 2011a; Greiff et al., 2012, 2015a; Alison et al., 2013; Gobert et al., 2015; Greiff and Fischer, 2013; Herde et al., 2016; Stadler et al., 2016), papers about *applications* (Fischer and Neubert, 2015; Ederer et al., 2016; Tremblay et al., 2017), papers about *differential effects* (Barth and Funke, 2010; Danner et al., 2011b; Beckmann and Goode, 2014; Greiff and Neubert, 2014; Scherer et al., 2015; Meißner et al., 2016; Wüstenberg et al., 2016), one paper about *developmental effects* (Frischkorn et al., 2014), one paper with a *neuroscience* background (Osman, 2012)<sup>1</sup>, papers about *cultural differences* (Güss and Dörner, 2011; Sonnleitner et al., 2014; Güss et al., 2015), papers about *validity issues* (Goode and Beckmann, 2010; Greiff et al., 2013c; Schweizer et al., 2013; Mainert et al., 2015; Funke et al., 2017; Greiff et al., 2017, 2015b; Kretzschmar et al., 2016; Kretzschmar, 2017), *review papers* and *meta-analyses* (Osman, 2010; Stadler et al., 2015), and finally *books* (Qudrat-Ullah, 2015; Csapó and Funke, 2017b) and *book chapters* (Funke, 2012; Hotaling et al., 2015; Funke and Greiff, 2017; Greiff and Funke, 2017; Csapó and Funke, 2017a; Fischer et al., 2017; Molnár et al., 2017; Tobinski and Fritz, 2017; Viehrig et al., 2017). In addition, a new “Journal of Dynamic Decision Making” (JDDM) has been launched (Fischer et al., 2015, 2016) to give the field an open-access outlet for research and discussion.

<sup>1</sup> The fMRI-paper from Anderson (2012) uses the term “complex problem solving” for tasks that do not fall in our understanding of CPS and is therefore excluded from this list.

This paper aims to clarify aspects of validity: what should be meant by the term CPS and what not? This clarification seems necessary because misunderstandings in recent publications provide – from our point of view – a potentially misleading picture of the construct. We start this article with a historical review before attempting to systematize different positions. We conclude with a working definition.

## HISTORICAL REVIEW

The concept behind CPS goes back to the German phrase “komplexes Problemlösen” (CPS; the term “komplexes Problemlösen” was used as a book title by Funke, 1986). The concept was introduced in Germany by Dörner and colleagues in the mid-1970s (see Dörner et al., 1975; Dörner, 1975) for the first time. The German phrase was later translated to CPS in the titles of two edited volumes by Sternberg and Frensch (1991) and Frensch and Funke (1995a) that collected papers from different research traditions. Even though it looks as though the term was coined in the 1970s, Edwards (1962) used the term “dynamic decision making” to describe decisions that come in a sequence. He compared static with dynamic decision making, writing:

In dynamic situations, a new complication not found in the static situations arises. The environment in which the decision is set may be changing, either as a function of the sequence of decisions, or independently of them, or both. It is this possibility of an environment which changes while you collect information about it which makes the task of dynamic decision theory so difficult and so much fun. (p. 60)

The ability to solve complex problems is typically measured via dynamic systems that contain several interrelated variables that participants need to alter. Early work (see, e.g., Dörner, 1980) used a simulation scenario called “Lohhausen” that contained more than 2000 variables that represented the activities of a small town: Participants had to take over the role of a mayor for a simulated period of 10 years. The simulation condensed these ten years to ten hours in real time. Later, researchers used smaller dynamic systems as scenarios either based on linear equations (see, e.g., Funke, 1993) or on finite state automata (see, e.g., Buchner and Funke, 1993). In these contexts, CPS consisted of the identification and control of dynamic task environments that were previously unknown to the participants. Different task environments came along with different degrees of fidelity (Gray, 2002).

According to Funke (2012), the typical attributes of complex systems are (a) *complexity* of the problem situation which is usually represented by the sheer number of involved variables; (b) *connectivity* and mutual dependencies between involved variables; (c) *dynamics* of the situation, which reflects the role of time and developments within a system; (d) *intransparency* (in part or full) about the involved variables and their current values; and (e) *polytely* (greek term for “many goals”), representing goal conflicts on different levels of analysis. This mixture of features is similar to what is called VUCA (volatility, uncertainty,



complexity, ambiguity) in modern approaches to management (e.g., Mack et al., 2016).

In his evaluation of the CPS movement, Sternberg (1995) compared (young) European approaches to CPS with (older) American research on expertise. His analysis of the differences between the European and American traditions shows advantages but also potential drawbacks for each side. He states (p. 301): “I believe that although there are problems with the European approach, it deals with some fundamental questions that American research scarcely addresses.” So, even though the echo of the European approach did not enjoy strong resonance in the US at that time, it was valued by scholars like Sternberg and others. Before attending to validity issues, we will first present a short review of different streams.

## DIFFERENT APPROACHES TO CPS

In the short history of CPS research, different approaches can be identified (Buchner, 1995; Fischer et al., 2017). To systematize, we differentiate between the following five lines of research:

- (a) The search for *individual differences* comprises studies identifying interindividual differences that affect the ability to solve complex problems. This line of research is reflected, for example, in the early work by Dörner et al. (1983) and their “Lohhausen” study. Here, naïve student participants took over the role of the mayor of a small simulated town named Lohhausen for a simulation period of ten years. According to the results of the authors, it is not intelligence (as measured by conventional IQ tests) that predicts performance, but it is the ability to stay calm in the face of a challenging situation and the ability to switch easily between an analytic mode of processing and a more holistic one.
- (b) The search for *cognitive processes* deals with the processes behind understanding complex dynamic systems. Representative of this line of research is, for example, Berry and Broadbent’s (1984) work on implicit and explicit learning processes when people interact with a dynamic system called “Sugar Production”. They found that those who perform best in controlling a dynamic system can do so implicitly, without explicit knowledge of details regarding the systems’ relations.
- (c) The search for *system factors* seeks to identify the aspects of dynamic systems that determine the difficulty of complex problems and make some problems harder than others. Representative of this line of research is, for example, work by Funke (1985), who systematically varied the number of causal effects within a dynamic system or the presence/absence of eigendynamics. He found, for example, that solution quality decreases as the number of systems relations increases.
- (d) The *psychometric approach* develops measurement instruments that can be used as an alternative to classical IQ tests, as something that goes “beyond IQ”. The MicroDYN approach (Wüstenberg et al., 2012) is representative for this
- line of research that presents an alternative to reasoning tests (like Raven matrices). These authors demonstrated that a small improvement in predicting school grade point average beyond reasoning is possible with MicroDYN tests.
- (e) The *experimental approach* explores CPS under different experimental conditions. This approach uses CPS assessment instruments to test hypotheses derived from psychological theories and is sometimes used in research about cognitive processes (see above). Exemplary for this line of research is the work by Rohe et al. (2016), who test the usefulness of “motto goals” in the context of complex problems compared to more traditional learning and performance goals. Motto goals differ from pure performance goals by activating positive affect and should lead to better goal attainment especially in complex situations (the mentioned study found no effect).

To be clear: these five approaches are not mutually exclusive and do overlap. But the differentiation helps to identify different research communities and different traditions. These communities had different opinions about scaling complexity.

## THE RACE FOR COMPLEXITY: USE OF MORE AND MORE COMPLEX SYSTEMS

In the early years of CPS research, microworlds started with systems containing about 20 variables (“Tailorshop”), soon reached 60 variables (“Moro”), and culminated in systems with about 2000 variables (“Lohhausen”). This race for complexity ended with the introduction of the concept of “minimal complex systems” (MCS; Greiff and Funke, 2009; Funke and Greiff, 2017), which ushered in a search for the lower bound of complexity instead of the higher bound, which could not be defined as easily. The idea behind this concept was that whereas the upper limits of complexity are unbound, the lower limits might be identifiable. Imagine starting with a simple system containing two variables with a simple linear connection between them; then, step by step, increase the number of variables and/or the type of connections. One soon reaches a point where the system can no longer be considered simple and has become a “complex system”. This point represents a minimal complex system. Despite some research having been conducted in this direction, the point of transition from simple to complex has not been identified clearly as of yet.

Some years later, the original “minimal complex systems” approach (Greiff and Funke, 2009) shifted to the “multiple complex systems” approach (Greiff et al., 2013a). This shift is more than a slight change in wording: it is important because it taps into the issue of validity directly. *Minimal* complex systems have been introduced in the context of challenges from large-scale assessments like PISA 2012 that measure new aspects of problem solving, namely *interactive* problems besides static problem solving (Greiff and Funke, 2017). PISA 2012 required test developers to remain within testing time constraints (given by the school class schedule). Also, test developers needed a large item pool for the construction of a broad class of problem

solving items. It was clear from the beginning that MCS deal with simple dynamic situations that require controlled interaction: the exploration and control of simple ticket machines, simple mobile phones, or simple MP3 players (all of these example domains were developed within PISA 2012) – rather than really *complex* situations like managerial or political decision making.

As a consequence of this subtle but important shift in interpreting the letters MCS, the definition of CPS became a subject of debate recently (Funke, 2014a; Greiff and Martin, 2014; Funke et al., 2017). In the words of Funke (2014b, p. 495):

It is funny that problems that nowadays come under the term ‘CPS’, are less complex (in terms of the previously described attributes of complex situations) than at the beginning of this new research tradition. The emphasis on psychometric qualities has led to a loss of variety. Systems thinking requires more than analyzing models with two or three linear equations – nonlinearity, cyclicity, rebound effects, etc. are inherent features of complex problems and should show up at least in some of the problems used for research and assessment purposes. Minimal complex systems run the danger of becoming minimal valid systems.

Searching for minimal complex systems is not the same as gaining insight into the way how humans deal with complexity and uncertainty. For psychometric purposes, it is appropriate to reduce complexity to a minimum; for understanding problem solving under conditions of overload, intransparency, and dynamics, it is necessary to realize those attributes with reasonable strength. This aspect is illustrated in the next section.

## IMPORTANCE OF THE VALIDITY ISSUE

The most important reason for discussing the question of what complex problem solving is and what it is not stems from its phenomenology: if we lose sight of our phenomena, we are no longer doing good psychology. The relevant phenomena in the context of complex problems encompass many important aspects. In this section, we discuss four phenomena that are specific to complex problems. We consider these phenomena as critical for theory development and for the construction of assessment instruments (i.e., microworlds). These phenomena require theories for explaining them and they require assessment instruments eliciting them in a reliable way.

The first phenomenon is the emergency reaction of the intellectual system (Dörner, 1980): When dealing with complex systems, actors tend to (a) reduce their intellectual level by decreasing self-reflections, by decreasing their intentions, by stereotyping, and by reducing their realization of intentions, (b) they show a tendency for fast action with increased readiness for risk, with increased violations of rules, and with increased tendency to escape the situation, and (c) they degenerate their hypotheses formation by construction of more global hypotheses and reduced tests of hypotheses, by increasing entrenchment, and by decontextualizing their goals. This phenomenon illustrates the strong *connection between cognition, emotion, and motivation*

that has been emphasized by Dörner (see, e.g., Dörner and Güss, 2013) from the beginning of his research tradition; the emergency reaction reveals a shift in the mode of information processing under the pressure of complexity.

The second phenomenon comprises cross-cultural differences with respect to strategy use (Strohschneider and Güss, 1999; Güss and Wiley, 2007; Güss et al., 2015). Results from complex task environments illustrate the strong influence of *context and background knowledge* to an extent that cannot be found for knowledge-poor problems. For example, in a comparison between Brazilian and German participants, it turned out that Brazilians accept the given problem descriptions and are more optimistic about the results of their efforts, whereas Germans tend to inquire more about the background of the problems and take a more active approach but are less optimistic (according to Strohschneider and Güss, 1998, p. 695).

The third phenomenon relates to failures that occur during the planning and acting stages (Jansson, 1994; Ramnarayan et al., 1997), illustrating that rational procedures seem to be unlikely to be used in complex situations. The *potential for failures* (Dörner, 1996) rises with the complexity of the problem. Jansson (1994) presents seven major areas for failures with complex situations: acting directly on current feedback; insufficient systematization; insufficient control of hypotheses and strategies; lack of self-reflection; selective information gathering; selective decision making; and thematic vagabonding.

The fourth phenomenon describes (a lack of) training and transfer effects (Kretzschmar and Süß, 2015), which again illustrates the *context dependency of strategies and knowledge* (i.e., there is no strategy that is so universal that it can be used in many different problem situations). In their own experiment, the authors could show training effects only for knowledge acquisition, not for knowledge application. Only with specific feedback, performance in complex environments can be increased (Engelhart et al., 2017).

These four phenomena illustrate why the type of complexity (or degree of simplicity) used in research really matters. Furthermore, they demonstrate effects that are specific for complex problems, but not for toy problems. These phenomena direct the attention to the important question: does the stimulus material used (i.e., the computer-simulated microworld) tap and elicit the manifold of phenomena described above?

Dealing with partly unknown complex systems requires courage, wisdom, knowledge, grit, and creativity. In creativity research, “little c” and “BIG C” are used to differentiate between everyday creativity and eminent creativity (Beghetto and Kaufman, 2007; Kaufman and Beghetto, 2009). Everyday creativity is important for solving everyday problems (e.g., finding a clever fix for a broken spoke on my bicycle), eminent creativity changes the world (e.g., inventing solar cells for energy production). Maybe problem solving research should use a similar differentiation between “little p” and “BIG P” to mark toy problems on the one side and big societal challenges on the other. The question then remains: what can we learn about BIG P by studying little p? What phenomena are present in both types, and what phenomena are unique to each of the two extremes?

## ON METHODS

Discussing research on CPS requires reflecting on the field's research methods. Even if the experimental approach has been successful for testing hypotheses (for an overview of older work, see Funke, 1995), other methods might provide additional and novel insights. Complex phenomena require complex approaches to understand them. The complex nature of complex systems imposes limitations on psychological experiments: The more complex the environments, the more difficult is it to keep conditions under experimental control. And if experiments have to be run in labs one should bring enough complexity into the lab to establish the phenomena mentioned, at least in part.

There are interesting options to be explored (again): *think-aloud protocols*, which have been discredited for many years (Nisbett and Wilson, 1977) and yet are a valuable source for theory testing (Ericsson and Simon, 1983); *introspection* (Jäkel and Schreiber, 2013), which seems to be banned from psychological methods but nevertheless offers insights into thought processes; the use of *life-streaming* (Wendt, 2017), a medium in which streamers generate a video stream of think-aloud data in computer-gaming; *political decision-making* (Dhami et al., 2015) that demonstrates error-proneness in groups; historical *case studies* (Dörner and Güss, 2011) that give insights into the thinking styles of political leaders; the use of the *critical incident* technique (Reuschenbach, 2008) to construct complex scenarios; and simulations with different degrees of *fidelity* (Gray, 2002).

The methods tool box is full of instruments that have to be explored more carefully before any individual instrument receives a ban or research narrows its focus to only one paradigm for data collection. Brehmer and Dörner (1993) discussed the tensions between “research in the laboratory and research in the field”, optimistically concluding “that the new methodology of computer-simulated microworlds will provide us with the means to bridge the gap between the laboratory and the field” (p. 183). The idea behind this optimism was that computer-simulated scenarios would bring more complexity from the outside world into the controlled lab environment. But this is not true for all simulated scenarios. In his paper on simulated environments, Gray (2002) differentiated computer-simulated environments with respect to three dimensions: (1) tractability (“the more training subjects require before they can use a simulated task environment, the less tractable it is”, p. 211), correspondence (“High correspondence simulated task environments simulate many aspects of one task environment. Low correspondence simulated task environments simulate one aspect of many task environments”, p. 214), and engagement (“A simulated task environment is engaging to the degree to which it involves and occupies the participants; that is, the degree to which they agree to take it seriously”, p. 217). But the mere fact that a task is called a “computer-simulated task environment” does not mean anything specific in terms of these three dimensions. This is one of several reasons why we should differentiate between those studies that do not address the core features of CPS and those that do.

## WHAT IS NOT CPS?

Even though a growing number of references claiming to deal with complex problems exist (e.g., Greiff and Wüstenberg, 2015; Greiff et al., 2016), it would be better to label the requirements within these tasks “dynamic problem solving,” as it has been done adequately in earlier work (Greiff et al., 2012). The dynamics behind on-off-switches (Thimbleby, 2007) are remarkable but not really complex. Small nonlinear systems that exhibit stunningly complex and unstable behavior do exist – but they are not used in psychometric assessments of so-called CPS. There are other small systems (like MicroDYN scenarios: Greiff and Wüstenberg, 2014) that exhibit simple forms of system behavior that are completely predictable and stable. This type of simple systems is used frequently. It is even offered commercially as a complex problem-solving test called COMPRO (Greiff and Wüstenberg, 2015) for business applications. But a closer look reveals that the label is not used correctly; within COMPRO, the used linear equations are far from being complex and the system can be handled properly by using only one strategy (see for more details Funke et al., 2017).

Why do simple linear systems not fall within CPS? At the surface, nonlinear and linear systems might appear similar because both only include 3–5 variables. But the difference is in terms of systems behavior as well as strategies and learning. If the behavior is simple (as in linear systems where more input is related to more output and vice versa), the system can be easily understood (participants in the MicroDYN world have 3 minutes to explore a complex system). If the behavior is complex (as in systems that contain strange attractors or negative feedback loops), things become more complicated and much more observation is needed to identify the hidden structure of the unknown system (Berry and Broadbent, 1984; Hundertmark et al., 2015).

Another issue is learning. If tasks can be solved using a single (and not so complicated) strategy, steep learning curves are to be expected. The shift from problem solving to learned routine behavior occurs rapidly, as was demonstrated by Luchins (1942). In his water jar experiments, participants quickly acquired a specific strategy (a mental set) for solving certain measurement problems that they later continued applying to problems that would have allowed for easier approaches. In the case of complex systems, learning can occur only on very general, abstract levels because it is difficult for human observers to make specific predictions. Routines dealing with complex systems are quite different from routines relating to linear systems.

What should not be studied under the label of CPS are pure learning effects, multiple-cue probability learning, or tasks that can be solved using a single strategy. This last issue is a problem for MicroDYN tasks that rely strongly on the VOTAT strategy (“vary one thing at a time”; see Tschirgi, 1980). In real-life, it is hard to imagine a business manager trying to solve her or his problems by means of VOTAT.



## WHAT IS CPS?

In the early days of CPS research, planet Earth's dynamics and complexities gained attention through such books as "The limits to growth" (Meadows et al., 1972) and "Beyond the limits" (Meadows et al., 1992). In the current decade, for example, the World Economic Forum (2016) attempts to identify the complexities and risks of our modern world. In order to understand the meaning of complexity and uncertainty, taking a look at the worlds' most pressing issues is helpful. Searching for strategies to cope with these problems is a difficult task: surely there is no place for the simple principle of "vary-one-thing-at-a-time" (VOTAT) when it comes to global problems. The VOTAT strategy is helpful in the context of simple problems (Wüstenberg et al., 2014); therefore, whether or not VOTAT is helpful in a given problem situation helps us distinguish simple from complex problems.

Because there exist no clear-cut strategies for complex problems, typical failures occur when dealing with uncertainty (Dörner, 1996; Güss et al., 2015). Ramnarayan et al. (1997) put together a list of generic errors (e.g., not developing adequate action plans; lack of background control; learning from experience blocked by stereotype knowledge; reactive instead of proactive action) that are typical of knowledge-rich complex systems but cannot be found in simple problems.

Complex problem solving is not a one-dimensional, low-level construct. On the contrary, CPS is a multi-dimensional bundle of competencies existing at a high level of abstraction, similar to intelligence (but going beyond IQ). As Funke et al. (2018) state: "Assessment of transversal (in educational contexts: cross-curricular) competencies cannot be done with one or two types of assessment. The plurality of skills and competencies requires a plurality of assessment instruments."

There are at least three different aspects of complex systems that are part of our understanding of a complex system: (1) a complex system can be described at different levels of abstraction; (2) a complex system develops over time, has a history, a current state, and a (potentially unpredictable) future; (3) a complex system is knowledge-rich and activates a large semantic network, together with a broad list of potential strategies (domain-specific as well as domain-general).

Complex problem solving is not only a cognitive process but is also an emotional one (Spering et al., 2005; Barth and Funke, 2010) and strongly dependent on motivation (low-stakes versus high-stakes testing; see Hermes and Stelling, 2016).

Furthermore, CPS is a dynamic process unfolding over time, with different phases and with more differentiation than simply knowledge acquisition and knowledge application. Ideally, the process should entail identifying problems (see Dillon, 1982; Lee and Cho, 2007), even if in experimental settings, problems are provided to participants *a priori*. The more complex and open a given situation, the more options can be generated (T. S. Schweizer et al., 2016). In closed problems, these processes do not occur in the same way.

In analogy to the difference between formative (process-oriented) and summative (result-oriented) assessment (William and Black, 1996; Bennett, 2011), CPS should not be reduced to the

mere outcome of a solution process. The process leading up to the solution, including detours and errors made along the way, might provide a more differentiated impression of a person's problem-solving abilities and competencies than the final result of such a process. This is one of the reasons why CPS environments are not, in fact, complex intelligence tests: research on CPS is not only about the outcome of the decision process, but it is also about the problem-solving process itself.

Complex problem solving is part of our daily life: finding the right person to share one's life with, choosing a career that not only makes money, but that also makes us happy. Of course, CPS is not restricted to personal problems – life on Earth gives us many hard nuts to crack: climate change, population growth, the threat of war, the use and distribution of natural resources. In sum, many societal challenges can be seen as complex problems. To reduce that complexity to a one-hour lab activity on a random Friday afternoon puts it out of context and does not address CPS issues.

Theories about CPS should specify which populations they apply to. Across populations, one thing to consider is prior knowledge. CPS research with experts (e.g., Dew et al., 2009) is quite different from problem solving research using tasks that intentionally do not require any specific prior knowledge (see, e.g., Beckmann and Goode, 2014).

More than 20 years ago, Frensch and Funke (1995b) 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. (p. 18)

The above definition is rather formal and does not account for content or relations between the simulation and the real world. In a sense, we need a new definition of CPS that addresses these issues. Based on our previous arguments, we propose the following working definition:

Complex problem solving is a collection of self-regulated psychological processes and activities necessary in dynamic environments to achieve ill-defined goals that cannot be reached by routine actions. Creative combinations of knowledge and a broad set of strategies are needed. Solutions are often more bricolage than perfect or optimal. The problem-solving process combines cognitive, emotional, and motivational aspects, particularly in high-stakes situations. Complex problems usually involve knowledge-rich requirements and collaboration among different persons.

The main differences to the older definition lie in the emphasis on (a) the self-regulation of processes, (b) creativity (as opposed to routine behavior), (c) the bricolage type of solution, and (d) the

role of high-stakes challenges. Our new definition incorporates some aspects that have been discussed in this review but were not reflected in the 1995 definition, which focused on attributes of complex problems like dynamics or intransparency.

This leads us to the final reflection about the role of CPS for dealing with uncertainty and complexity in real life. We will distinguish thinking from reasoning and introduce the sense of possibility as an important aspect of validity.

## CPS AS COMBINING REASONING AND THINKING IN AN UNCERTAIN REALITY

Leading up to the Battle of Borodino in Leo Tolstoy's novel "War and Peace", Prince Andrei Bolkonsky explains the concept of war to his friend Pierre. Pierre expects war to resemble a game of chess: You position the troops and attempt to defeat your opponent by moving them in different directions.

"Far from it!", Andrei responds. "In chess, you know the knight and his moves, you know the pawn and his combat strength. While in war, a battalion is sometimes stronger than a division and sometimes weaker than a company; it all depends on circumstances that can never be known. In war, you do not know the position of your enemy; some things you might be able to observe, some things you have to divine (but that depends on your ability to do so!) and many things cannot even be guessed at. In chess, you can see all of your opponent's possible moves. In war, that is impossible. If you decide to attack, you cannot know whether the necessary conditions are met for you to succeed. Many a time, you cannot even know whether your troops will follow your orders..."

In essence, war is characterized by a high degree of uncertainty. A good commander (or politician) can add to that what he or she sees, tentatively fill in the blanks – and not just by means of logical deduction but also by intelligently bridging missing links. A bad commander extrapolates from what he sees and thus arrives at improper conclusions.

Many languages differentiate between two modes of mentalizing; for instance, the English language distinguishes between 'thinking' and 'reasoning'. Reasoning denotes acute and exact mentalizing involving logical deductions. Such deductions are usually based on evidence and counterevidence. Thinking, however, is what is required to write novels. It is the construction of an initially unknown reality. But it is not a pipe dream, an unfounded process of fabrication. Rather, thinking asks us to imagine reality ("Wirklichkeitsfantasie"). In other words, a novelist has to possess a "sense of possibility" ("Möglichkeitssinn", Robert Musil; in German, sense of possibility is often used synonymously with imagination even though imagination is not the same as sense of possibility, for imagination also encapsulates the impossible). This sense of possibility entails knowing the whole (or several wholes) or being able to construe an unknown whole that could accommodate a known part. The whole has to align with sociological and geographical givens, with the mentality of certain peoples or groups, and with the laws of physics and chemistry. Otherwise, the entire venture is ill-founded. A sense of possibility does not

aim for the moon but imagines something that might be possible but has not been considered possible or even potentially possible so far.

Thinking is a means to eliminate uncertainty. This process requires both of the modes of thinking we have discussed thus far. Economic, political, or ecological decisions require us to first consider the situation at hand. Though certain situational aspects can be known, but many cannot. In fact, von Clausewitz (1832) posits that only about 25% of the necessary information is available when a military decision needs to be made. Even then, there is no way to guarantee that whatever information is available is also correct: Even if a piece of information was completely accurate yesterday, it might no longer apply today.

Once our sense of possibility has helped grasping a situation, problem solvers need to call on their reasoning skills. Not every situation requires the same action, and we may want to act this way or another to reach this or that goal. This appears logical, but it is a logic based on constantly shifting grounds: We cannot know whether necessary conditions are met, sometimes the assumptions we have made later turn out to be incorrect, and sometimes we have to revise our assumptions or make completely new ones. It is necessary to constantly switch between our sense of possibility and our sense of reality, that is, to switch between thinking and reasoning. It is an arduous process, and some people handle it well, while others do not.

If we are to believe Tuchman's (1984) book, "The March of Folly", most politicians and commanders are fools. According to Tuchman, not much has changed in the 3300 years that have elapsed since the misguided Trojans decided to welcome the left-behind wooden horse into their city that would end up dismantling Troy's defensive walls. The Trojans, too, had been warned, but decided not to heed the warning. Although Laocoön had revealed the horse's true nature to them by attacking it with a spear, making the weapons inside the horse ring, the Trojans refused to see the forest for the trees. They did not want to listen, they wanted the war to be over, and this desire ended up shaping their perception.

The objective of psychology is to predict and explain human actions and behavior as accurately as possible. However, thinking cannot be investigated by limiting its study to neatly confined fractions of reality such as the realms of propositional logic, chess, Go tasks, the Tower of Hanoi, and so forth. Within these systems, there is little need for a sense of possibility. But a sense of possibility – the ability to divine and construe an unknown reality – is at least as important as logical reasoning skills. Not researching the sense of possibility limits the validity of psychological research. All economic and political decision making draws upon this sense of possibility. By not exploring it, psychological research dedicated to the study of thinking cannot further the understanding of politicians' competence and the reasons that underlie political mistakes. Christopher Clark identifies European diplomats', politicians', and commanders' inability to form an accurate representation of reality as a reason for the outbreak of World War I. According to Clark's (2012) book, "The Sleepwalkers", the politicians of the time lived in their own make-believe world, wrongfully assuming that it was the same world everyone else inhabited. If CPS research wants to

make significant contributions to the world, it has to acknowledge complexity and uncertainty as important aspects of it.

## CONCLUSION

For more than 40 years, CPS has been a new subject of psychological research. During this time period, the initial emphasis on analyzing how humans deal with complex, dynamic, and uncertain situations has been lost. What is subsumed under the heading of CPS in modern research has lost the original complexities of real-life problems. From our point of view, the challenges of the 21st century require a return to the origins of this research tradition. We would encourage researchers in the field of problem solving to come back to the original ideas. There is enough complexity and uncertainty in the world to be studied. Improving our understanding of how humans deal with these global and pressing problems would be a worthwhile enterprise.

## AUTHOR CONTRIBUTIONS

JF drafted a first version of the manuscript, DD added further text and commented on the draft. JF finalized the manuscript.

## REFERENCES

- Alison, L., van den Heuvel, C., Waring, S., Power, N., Long, A., O'Hara, T., et al. (2013). Immersive simulated learning environments for researching critical incidents: a knowledge synthesis of the literature and experiences of studying high-risk strategic decision making. *J. Cogn. Eng. Deci. Mak.* 7, 255–272. doi: 10.1177/1555343412468113
- Anderson, J. R. (2012). Tracking problem solving by multivariate pattern analysis and hidden markov model algorithms. *Neuropsychologia* 50, 487–498. doi: 10.1016/j.neuropsychologia.2011.07.025
- Barth, C. M., and Funke, J. (2010). Negative affective environments improve complex solving performance. *Cogn. Emot.* 24, 1259–1268. doi: 10.1080/02699930903223766
- Beckmann, J. F., and Goode, N. (2014). The benefit of being naïve and knowing it: the unfavourable impact of perceived context familiarity on learning in complex problem solving tasks. *Instruct. Sci.* 42, 271–290. doi: 10.1007/s11251-013-9280-7
- Beghetto, R. A., and Kaufman, J. C. (2007). Toward a broader conception of creativity: a case for “mini-c” creativity. *Psychol. Aesthetics Creat. Arts* 1, 73–79. doi: 10.1037/1931-3896.1.2.73
- Bennett, R. E. (2011). Formative assessment: a critical review. *Assess. Educ. Princ. Policy Pract.* 18, 5–25. doi: 10.1080/0969594X.2010.513678
- Berry, D. C., and Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *Q. J. Exp. Psychol.* 36, 209–231. doi: 10.1080/14640748408402156
- Bleche, C., and Funke, J. (2010). You cannot have your cake and eat it, too: how induced goal conflicts affect complex problem solving. *Open Psychol. J.* 3, 42–53. doi: 10.2174/1874350101003010042
- Brehmer, B., and Dörner, D. (1993). Experiments with computer-simulated microworlds: escaping both the narrow straits of the laboratory and the deep blue sea of the field study. *Comput. Hum. Behav.* 9, 171–184. doi: 10.1016/0747-5632(93)90005-D
- Buchner, A. (1995). “Basic topics and approaches to the study of complex problem solving,” in *Complex Problem Solving: The European Perspective*, eds P. A. Frensch and J. Funke (Hillsdale, NJ: Erlbaum), 27–63.
- Buchner, A., and Funke, J. (1993). Finite state automata: dynamic task environments in problem solving research. *Q. J. Exp. Psychol.* 46A, 83–118. doi: 10.1080/14640749308401068
- Clark, C. (2012). *The Sleepwalkers: How Europe Went to War in 1914*. London: Allen Lane.
- Csapó, B., and Funke, J. (2017a). “The development and assessment of problem solving in 21st-century schools,” in *The Nature of Problem Solving: Using Research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (Paris: OECD Publishing), 19–31.
- Csapó, B., and Funke, J. (eds) (2017b). *The Nature of Problem Solving. Using Research to Inspire 21st Century Learning*. Paris: OECD Publishing.
- Danner, D., Hagemann, D., Holt, D. V., Hager, M., Schankin, A., Wüstenberg, S., et al. (2011a). Measuring performance in dynamic decision making. Reliability and validity of the Tailorshop simulation. *J. Ind. Differ.* 32, 225–233. doi: 10.1027/1614-0001/a000055
- Danner, D., Hagemann, D., Schankin, A., Hager, M., and Funke, J. (2011b). Beyond IQ: a latent state-trait analysis of general intelligence, dynamic decision making, and implicit learning. *Intelligence* 39, 323–334. doi: 10.1016/j.intell.2011.06.004
- Dew, N., Read, S., Sarasvathy, S. D., and Wiltbank, R. (2009). Effectual versus predictive logics in entrepreneurial decision-making: differences between experts and novices. *J. Bus. Ventur.* 24, 287–309. doi: 10.1016/j.jbusvent.2008.02.002
- Dhami, M. K., Mandel, D. R., Mellers, B. A., and Tetlock, P. E. (2015). Improving intelligence analysis with decision science. *Perspect. Psychol. Sci.* 10, 753–757. doi: 10.1177/1745691615598511
- Dillon, J. T. (1982). Problem finding and solving. *J. Creat. Behav.* 16, 97–111. doi: 10.1002/j.2162-6057.1982.tb00326.x
- Dörner, D. (1975). Wie Menschen eine Welt verbessern wollten [How people wanted to improve a world]. *Bild Der Wissenschaft* 12, 48–53.
- Dörner, D. (1980). On the difficulties people have in dealing with complexity. *Simulat. Gam.* 11, 87–106. doi: 10.1177/104687818001100108
- Dörner, D. (1996). *The Logic of Failure: Recognizing and Avoiding Error in Complex Situations*. New York, NY: Basic Books.
- Dörner, D., Drewes, U., and Reither, F. (1975). “Über das Problemlösen in sehr komplexen Realitätsbereichen,” in *Bericht über den 29. Kongress der DGfPs in Salzburg 1974, Band 1*, ed. W. H. Tack (Göttingen: Hogrefe), 339–340.
- Dörner, D., and Güss, C. D. (2011). A psychological analysis of Adolf Hitler's decision making as commander in chief: summa confidentia et nimium metus. *Rev. Gen. Psychol.* 15, 37–49. doi: 10.1037/a0022375
- Dörner, D., and Güss, C. D. (2013). PSI: a computational architecture of cognition, motivation, and emotion. *Rev. Gen. Psychol.* 17, 297–317. doi: 10.1037/a0032947

## AUTHORS NOTE

After more than 40 years of controversial discussions between both authors, this is the first joint paper. We are happy to have done this now! We have found common ground!

## ACKNOWLEDGMENTS

The authors thank the Deutsche Forschungsgemeinschaft (DFG) for the continuous support of their research over many years. Thanks to Daniel Holt for his comments on validity issues, thanks to Julia Nolte who helped us by translating German text excerpts into readable English and helped us, together with Keri Hartman, to improve our style and grammar – thanks for that! We also thank the two reviewers for their helpful critical comments on earlier versions of this manuscript. Finally, we acknowledge financial support by Deutsche Forschungsgemeinschaft and Ruprecht-Karls-Universität Heidelberg within their funding programme *Open Access Publishing*.



- Dörner, D., Kreuzig, H. W., Reither, F., and Stäudel, T. (1983). *Lohhausen. Vom Umgang mit Unbestimmtheit und Komplexität*. Bern: Huber.
- Ederer, P., Patt, A., and Greiff, S. (2016). Complex problem-solving skills and innovativeness – evidence from occupational testing and regional data. *Eur. J. Educ.* 51, 244–256. doi: 10.1111/ejed.12176
- Edwards, W. (1962). Dynamic decision theory and probabilistic information processing. *Hum. Factors* 4, 59–73. doi: 10.1177/001872086200400201
- Engelhart, M., Funke, J., and Sager, S. (2017). A web-based feedback study on optimization-based training and analysis of human decision making. *J. Dynamic Dec. Mak.* 3, 1–23.
- Ericsson, K. A., and Simon, H. A. (1983). *Protocol Analysis: Verbal Reports As Data*. Cambridge, MA: Bradford.
- Fischer, A., Greiff, S., and Funke, J. (2017). “The history of complex problem solving,” in *The Nature of Problem Solving: Using Research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (Paris: OECD Publishing), 107–121.
- Fischer, A., Holt, D. V., and Funke, J. (2015). Promoting the growing field of dynamic decision making. *J. Dynamic Decis. Mak.* 1, 1–3. doi: 10.11588/jddm.2015.1.23807
- Fischer, A., Holt, D. V., and Funke, J. (2016). The first year of the “journal of dynamic decision making.” *J. Dynamic Decis. Mak.* 2, 1–2. doi: 10.11588/jddm.2016.1.28995
- Fischer, A., and Neubert, J. C. (2015). The multiple faces of complex problems: a model of problem solving competency and its implications for training and assessment. *J. Dynamic Decis. Mak.* 1, 1–14. doi: 10.11588/jddm.2015.1.23945
- Frensch, P. A., and Funke, J. (eds) (1995a). *Complex Problem Solving: The European Perspective*. Hillsdale, NJ: Erlbaum.
- Frensch, P. A., and Funke, J. (1995b). “Definitions, traditions, and a general framework for understanding complex problem solving,” in *Complex Problem Solving: The European Perspective*, eds P. A. Frensch and J. Funke (Hillsdale, NJ: Lawrence Erlbaum), 3–25.
- Frischkorn, G. T., Greiff, S., and Wüstenberg, S. (2014). The development of complex problem solving in adolescence: a latent growth curve analysis. *J. Educ. Psychol.* 106, 1004–1020. doi: 10.1037/a0037114
- Funke, J. (1985). Steuerung dynamischer Systeme durch Aufbau und Anwendung subjektiver Kausalmodelle. *Z. Psychol.* 193, 435–457.
- Funke, J. (1986). *Komplexes Problemlösen - Bestandsaufnahme und Perspektiven [Complex Problem Solving: Survey and Perspectives]*. Heidelberg: Springer.
- Funke, J. (1993). “Microworlds based on linear equation systems: a new approach to complex problem solving and experimental results,” in *The Cognitive Psychology of Knowledge*, eds G. Strube and K.-F. Wender (Amsterdam: Elsevier Science Publishers), 313–330.
- Funke, J. (1995). “Experimental research on complex problem solving,” in *Complex Problem Solving: The European Perspective*, eds P. A. Frensch and J. Funke (Hillsdale, NJ: Erlbaum), 243–268.
- Funke, J. (2010). Complex problem solving: a case for complex cognition? *Cogn. Process.* 11, 133–142. doi: 10.1007/s10339-009-0345-0
- Funke, J. (2012). “Complex problem solving,” in *Encyclopedia of the Sciences of Learning*, Vol. 38, ed. N. M. Seel (Heidelberg: Springer), 682–685.
- Funke, J. (2014a). Analysis of minimal complex systems and complex problem solving require different forms of causal cognition. *Front. Psychol.* 5:739. doi: 10.3389/fpsyg.2014.00739
- Funke, J. (2014b). “Problem solving: what are the important questions?” in *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, eds P. Bello, M. Guarini, M. McShane, and B. Scassellati (Austin, TX: Cognitive Science Society), 493–498.
- Funke, J., Fischer, A., and Holt, D. V. (2017). When less is less: solving multiple simple problems is not complex problem solving—A comment on Greiff et al. (2015). *J. Intell.* 5:5. doi: 10.3390/jintelligence5010005
- Funke, J., Fischer, A., and Holt, D. V. (2018). “Competencies for complexity: problem solving in the 21st century,” in *Assessment and Teaching of 21st Century Skills*, eds E. Care, P. Griffin, and M. Wilson (Dordrecht: Springer), 3.
- Funke, J., and Greiff, S. (2017). “Dynamic problem solving: multiple-item testing based on minimally complex systems,” in *Competence Assessment in Education. Research, Models and Instruments*, eds D. Leutner, J. Fleischer, J. Grünkorn, and E. Klieme (Heidelberg: Springer), 427–443.
- Gobert, J. D., Kim, Y. J., Pedro, M. A. S., Kennedy, M., and Betts, C. G. (2015). Using educational data mining to assess students’ skills at designing and conducting experiments within a complex systems microworld. *Think. Skills Creat.* 18, 81–90. doi: 10.1016/j.tsc.2015.04.008
- Goode, N., and Beckmann, J. F. (2010). You need to know: there is a causal relationship between structural knowledge and control performance in complex problem solving tasks. *Intelligence* 38, 345–352. doi: 10.1016/j.intell.2010.01.001
- Gray, W. D. (2002). Simulated task environments: the role of high-fidelity simulations, scaled worlds, synthetic environments, and laboratory tasks in basic and applied cognitive research. *Cogn. Sci. Q.* 2, 205–227.
- Greiff, S., and Fischer, A. (2013). Measuring complex problem solving: an educational application of psychological theories. *J. Educ. Res.* 5, 38–58.
- Greiff, S., Fischer, A., Stadler, M., and Wüstenberg, S. (2015a). Assessing complex problem-solving skills with multiple complex systems. *Think. Reason.* 21, 356–382. doi: 10.1080/13546783.2014.989263
- Greiff, S., Stadler, M., Sonnleitner, P., Wolff, C., and Martin, R. (2015b). Sometimes less is more: comparing the validity of complex problem solving measures. *Intelligence* 50, 100–113. doi: 10.1016/j.intell.2015.02.007
- Greiff, S., Fischer, A., Wüstenberg, S., Sonnleitner, P., Brunner, M., and Martin, R. (2013a). A multitrait-multimethod study of assessment instruments for complex problem solving. *Intelligence* 41, 579–596. doi: 10.1016/j.intell.2013.07.012
- Greiff, S., Holt, D. V., and Funke, J. (2013b). Perspectives on problem solving in educational assessment: analytical, interactive, and collaborative problem solving. *J. Problem Solv.* 5, 71–91. doi: 10.7771/1932-6246.1153
- Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., and Csapó, B. (2013c). Complex problem solving in educational contexts—something beyond g: concept, assessment, measurement invariance, and construct validity. *J. Educ. Psychol.* 105, 364–379. doi: 10.1037/a0031856
- Greiff, S., and Funke, J. (2009). “Measuring complex problem solving: the MicroDYN approach,” in *The Transition to Computer-Based Assessment. New Approaches to Skills Assessment and Implications for Large-Scale Testing*, eds F. Scheuermann and J. Björnsson (Luxembourg: Office for Official Publications of the European Communities), 157–163.
- Greiff, S., and Funke, J. (2017). “Interactive problem solving: exploring the potential of minimal complex systems,” in *The Nature of Problem Solving: Using Research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (Paris: OECD Publishing), 93–105.
- Greiff, S., and Martin, R. (2014). What you see is what you (don’t) get: a comment on Funke’s (2014) opinion paper. *Front. Psychol.* 5:1120. doi: 10.3389/fpsyg.2014.01120
- Greiff, S., and Neubert, J. C. (2014). On the relation of complex problem solving, personality, fluid intelligence, and academic achievement. *Learn. Ind. Diff.* 36, 37–48. doi: 10.1016/j.lindif.2014.08.003
- Greiff, S., Niepel, C., Scherer, R., and Martin, R. (2016). Understanding students’ performance in a computer-based assessment of complex problem solving: an analysis of behavioral data from computer-generated log files. *Comput. Hum. Behav.* 61, 36–46. doi: 10.1016/j.chb.2016.02.095
- Greiff, S., Stadler, M., Sonnleitner, P., Wolff, C., and Martin, R. (2017). Sometimes more is too much: a rejoinder to the commentaries on Greiff et al. (2015). *J. Intell.* 5:6. doi: 10.3390/jintelligence5010006
- Greiff, S., and Wüstenberg, S. (2014). Assessment with microworlds using MicroDYN: measurement invariance and latent mean comparisons. *Eur. J. Psychol. Assess.* 1, 1–11. doi: 10.1027/1015-5759/a000194
- Greiff, S., and Wüstenberg, S. (2015). Komplexer Problemlösetest COMPRO [Complex Problem-Solving Test COMPRO]. Mödling: Schuhfried.
- Greiff, S., Wüstenberg, S., and Funke, J. (2012). Dynamic problem solving: a new assessment perspective. *Appl. Psychol. Measure.* 36, 189–213. doi: 10.1177/0146621612439620
- Griffin, P., and Care, E. (2015). “The ATC21S method,” in *Assessment and Teaching of 21st Century Skills*, eds P. Griffin and E. Care (Dordrecht, NL: Springer), 3–33.
- Güss, C. D., and Dörner, D. (2011). Cultural differences in dynamic decision-making strategies in a non-linear, time-delayed task. *Cogn. Syst. Res.* 12, 365–376. doi: 10.1016/j.cogsys.2010.12.003
- Güss, C. D., Tuason, M. T., and Orduña, L. V. (2015). Strategies, tactics, and errors in dynamic decision making in an Asian sample. *J. Dynamic Decis. Mak.* 1, 1–14. doi: 10.11588/jddm.2015.1.13131

- Güss, C. D., and Wiley, B. (2007). Metacognition of problem-solving strategies in Brazil, India, and the United States. *J. Cogn. Cult.* 7, 1–25. doi: 10.1163/156853707X171793
- Herde, C. N., Wüstenberg, S., and Greiff, S. (2016). Assessment of complex problem solving: what we know and what we don't know. *Appl. Meas. Educ.* 29, 265–277. doi: 10.1080/08957347.2016.1209208
- Hermes, M., and Stelling, D. (2016). Context matters, but how much? Latent state – trait analysis of cognitive ability assessments. *Int. J. Sel. Assess.* 24, 285–295. doi: 10.1111/ijss.12147
- Hotaling, J. M., Fakhari, P., and Busemeyer, J. R. (2015). “Dynamic decision making,” in *International Encyclopedia of the Social & Behavioral Sciences*, 2nd Edn, eds N. J. Smelser and P. B. Bates (New York, NY: Elsevier), 709–714.
- Hundertmark, J., Holt, D. V., Fischer, A., Said, N., and Fischer, H. (2015). System structure and cognitive ability as predictors of performance in dynamic system control tasks. *J. Dynamic Deci. Mak.* 1, 1–10. doi: 10.11588/jddm.2015.1.26416
- Jäkel, F., and Schreiber, C. (2013). Introspection in problem solving. *J. Problem Solv.* 6, 20–33. doi: 10.7771/1932-6246.1131
- Jansson, A. (1994). Pathologies in dynamic decision making: consequences or precursors of failure? *Sprache Kogn.* 13, 160–173.
- Kaufman, J. C., and Beghetto, R. A. (2009). Beyond big and little: the four c model of creativity. *Rev. Gen. Psychol.* 13, 1–12. doi: 10.1037/a0013688
- Knauff, M., and Wolf, A. G. (2010). Complex cognition: the science of human reasoning, problem-solving, and decision-making. *Cogn. Process.* 11, 99–102. doi: 10.1007/s10339-010-0362-z
- Kretschmar, A. (2017). Sometimes less is not enough: a commentary on Greiff et al. (2015). *J. Intell.* 5:4. doi: 10.3390/jintelligence5010004
- Kretschmar, A., Neubert, J. C., Wüstenberg, S., and Greiff, S. (2016). Construct validity of complex problem solving: a comprehensive view on different facets of intelligence and school grades. *Intelligence* 54, 55–69. doi: 10.1016/j.intell.2015.11.004
- Kretschmar, A., and Süß, H.-M. (2015). A study on the training of complex problem solving competence. *J. Dynamic Deci. Mak.* 1, 1–14. doi: 10.11588/jddm.2015.1.15455
- Lee, H., and Cho, Y. (2007). Factors affecting problem finding depending on degree of structure of problem situation. *J. Educ. Res.* 101, 113–123. doi: 10.3200/JOER.101.2.113-125
- Leutner, D., Fleischer, J., Wirth, J., Greiff, S., and Funke, J. (2012). Analytische und dynamische Problemlösekompetenz im Lichte internationaler Schulleistungsvergleichsstudien: Untersuchungen zur Dimensionalität. *Psychol. Rundschau* 63, 34–42. doi: 10.1026/0033-3042/a000108
- Luchins, A. S. (1942). Mechanization in problem solving: the effect of Einstellung. *Psychol. Monogr.* 54, 1–95. doi: 10.1037/h0093502
- Mack, O., Khare, A., Krämer, A., and Burgartz, T. (eds) (2016). *Managing in a VUCA world*. Heidelberg: Springer.
- Mainert, J., Kretschmar, A., Neubert, J. C., and Greiff, S. (2015). Linking complex problem solving and general mental ability to career advancement: does a transversal skill reveal incremental predictive validity? *Int. J. Lifelong Educ.* 34, 393–411. doi: 10.1080/02601370.2015.1060024
- Mainzer, K. (2009). Challenges of complexity in the 21st century. An interdisciplinary introduction. *Eur. Rev.* 17, 219–236. doi: 10.1017/S1062798709000714
- Meadows, D. H., Meadows, D. L., and Randers, J. (1992). *Beyond the Limits*. Vermont, VA: Chelsea Green Publishing.
- Meadows, D. H., Meadows, D. L., Randers, J., and Behrens, W. W. (1972). *The Limits to Growth*. New York, NY: Universe Books.
- Meißner, A., Greiff, S., Frischkorn, G. T., and Steinmayr, R. (2016). Predicting complex problem solving and school grades with working memory and ability self-concept. *Learn. Ind. Differ.* 49, 323–331. doi: 10.1016/j.lindif.2016.04.006
- Molnár, G., Greiff, S., Wüstenberg, S., and Fischer, A. (2017). “Empirical study of computer-based assessment of domain-general complex problem-solving skills,” in *The Nature of Problem Solving: Using research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (Paris: OECD Publishing), 125–141.
- National Research Council (2011). *Assessing 21st Century Skills: Summary of a Workshop*. Washington, DC: The National Academies Press.
- Newell, A., Shaw, J. C., and Simon, H. A. (1959). A general problem-solving program for a computer. *Comput. Automat.* 8, 10–16.
- Nisbett, R. E., and Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychol. Rev.* 84, 231–259. doi: 10.1037/0033-295X.84.3.231
- OECD (2014). “PISA 2012 results,” in *Creative Problem Solving: Students' Skills in Tackling Real-Life problems*, Vol. 5 (Paris: OECD Publishing).
- Osman, M. (2010). Controlling uncertainty: a review of human behavior in complex dynamic environments. *Psychol. Bull.* 136, 65–86. doi: 10.1037/a0017815
- Osman, M. (2012). The role of reward in dynamic decision making. *Front. Neurosci.* 6:35. doi: 10.3389/fnins.2012.00035
- Qudrat-Ullah, H. (2015). *Better Decision Making in Complex, Dynamic Tasks. Training with Human-Facilitated Interactive Learning Environments*. Heidelberg: Springer.
- Ramnarayan, S., Strohschneider, S., and Schaub, H. (1997). Trappings of expertise and the pursuit of failure. *Simulat. Gam.* 28, 28–43. doi: 10.1177/1046878197281004
- Reuschenbach, B. (2008). *Planen und Problemlösen im Komplexen Handlungsfeld Pflege*. Berlin: Logos.
- Rohe, M., Funke, J., Storch, M., and Weber, J. (2016). Can motto goals outperform learning and performance goals? Influence of goal setting on performance, intrinsic motivation, processing style, and affect in a complex problem solving task. *J. Dynamic Deci. Mak.* 2, 1–15. doi: 10.11588/jddm.2016.1.28510
- Scherer, R., Greiff, S., and Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence* 48, 37–50. doi: 10.1016/j.intell.2014.10.003
- Schoppek, W., and Fischer, A. (2015). Complex problem solving – single ability or complex phenomenon? *Front. Psychol.* 6:1669. doi: 10.3389/fpsyg.2015.01669
- Schraw, G., Dunkle, M., and Bendixen, L. D. (1995). Cognitive processes in well-defined and ill-defined problem solving. *Appl. Cogn. Psychol.* 9, 523–538. doi: 10.1002/acp.2350090605
- Schweizer, F., Wüstenberg, S., and Greiff, S. (2013). Validity of the MicroDYN approach: complex problem solving predicts school grades beyond working memory capacity. *Learn. Ind. Differ.* 24, 42–52. doi: 10.1016/j.lindif.2012.12.011
- Schweizer, T. S., Schmalenberger, K. M., Eisenlohr-Moul, T. A., Mojzisch, A., Kaiser, S., and Funke, J. (2016). Cognitive and affective aspects of creative option generation in everyday life situations. *Front. Psychol.* 7:1132. doi: 10.3389/fpsyg.2016.01132
- Selten, R., Pittnauer, S., and Hohnisch, M. (2012). Dealing with dynamic decision problems when knowledge of the environment is limited: an approach based on goal systems. *J. Behav. Deci. Mak.* 25, 443–457. doi: 10.1002/bdm.738
- Simon, H. A. (1957). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations*, 2nd Edn. New York, NY: Macmillan.
- Sonnleitner, P., Brunner, M., Keller, U., and Martin, R. (2014). Differential relations between facets of complex problem solving and students' immigration background. *J. Educ. Psychol.* 106, 681–695. doi: 10.1037/a0035506
- Spering, M., Wagnen, D., and Funke, J. (2005). The role of emotions in complex problem solving. *Cogn. Emot.* 19, 1252–1261. doi: 10.1080/02699930500304886
- Stadler, M., Becker, N., Gödker, M., Leutner, D., and Greiff, S. (2015). Complex problem solving and intelligence: a meta-analysis. *Intelligence* 53, 92–101. doi: 10.1016/j.intell.2015.09.005
- Stadler, M., Niepel, C., and Greiff, S. (2016). Easily too difficult: estimating item difficulty in computer simulated microworlds. *Comput. Hum. Behav.* 65, 100–106. doi: 10.1016/j.chb.2016.08.025
- Sternberg, R. J. (1995). “Expertise in complex problem solving: a comparison of alternative conceptions,” in *Complex Problem Solving: The European Perspective*, eds P. A. Frensch and J. Funke (Hillsdale, NJ: Erlbaum), 295–321.
- Sternberg, R. J., and Frensch, P. A. (1991). *Complex Problem Solving: Principles and Mechanisms*. (eds) R. J. Sternberg and P. A. Frensch. Hillsdale, NJ: Erlbaum.
- Strohschneider, S., and Güss, C. D. (1998). Planning and problem solving: differences between Brazilian and German students. *J. Cross-Cult. Psychol.* 29, 695–716. doi: 10.1177/0022022198296002
- Strohschneider, S., and Güss, C. D. (1999). The fate of the Moros: a cross-cultural exploration of strategies in complex and dynamic decision making. *Int. J. Psychol.* 34, 235–252. doi: 10.1080/002075999399873
- Thimbleby, H. (2007). *Press On. Principles of Interaction*. Cambridge, MA: MIT Press.

- Tobinski, D. A., and Fritz, A. (2017). "EcoSphere: a new paradigm for problem solving in complex systems," in *The Nature of Problem Solving: Using Research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (Paris: OECD Publishing), 211–222.
- Tremblay, S., Gagnon, J.-F., Lafond, D., Hodgetts, H. M., Doiron, M., and Jeuniaux, P. P. J. M. H. (2017). A cognitive prosthesis for complex decision-making. *Appl. Ergon.* 58, 349–360. doi: 10.1016/j.apergo.2016.07.009
- Tschirgi, J. E. (1980). Sensible reasoning: a hypothesis about hypotheses. *Child Dev.* 51, 1–10. doi: 10.2307/1129583
- Tuchman, B. W. (1984). *The March of Folly. From Troy to Vietnam*. New York, NY: Ballantine Books.
- Verweij, M., and Thompson, M. (eds) (2006). *Clumsy Solutions for A Complex World. Governance, Politics and Plural Perceptions*. New York, NY: Palgrave Macmillan. doi: 10.1057/9780230624887
- Viehrieg, K., Siegmund, A., Funke, J., Wüstenberg, S., and Greiff, S. (2017). "The heidelberg inventory of geographic system competency model," in *Competence Assessment in Education. Research, Models and Instruments*, eds D. Leutner, J. Fleischer, J. Grünkorn, and E. Klieme (Heidelberg: Springer), 31–53.
- von Clausewitz, C. (1832). *Vom Kriege [On war]*. Berlin: Dämmler.
- Wendt, A. N. (2017). The empirical potential of live streaming beyond cognitive psychology. *J. Dynamic Deci. Mak.* 3, 1–9. doi: 10.11588/jddm.2017.1.33724
- William, D., and Black, P. (1996). Meanings and consequences: a basis for distinguishing formative and summative functions of assessment? *Br. Educ. Res. J.* 22, 537–548. doi: 10.1080/0141192960220502
- World Economic Forum (2015). *New Vision for Education Unlocking the Potential of Technology*. Geneva: World Economic Forum.
- World Economic Forum (2016). *Global Risks 2016: Insight Report*, 11th Edn. Geneva: World Economic Forum.
- Wüstenberg, S., Greiff, S., and Funke, J. (2012). Complex problem solving — more than reasoning? *Intelligence* 40, 1–14. doi: 10.1016/j.intell.2011.11.003
- Wüstenberg, S., Greiff, S., Vainikainen, M.-P., and Murphy, K. (2016). Individual differences in students' complex problem solving skills: how they evolve and what they imply. *J. Educ. Psychol.* 108, 1028–1044. doi: 10.1037/edu0000101
- Wüstenberg, S., Stadler, M., Hautamäki, J., and Greiff, S. (2014). The role of strategy knowledge for the application of strategies in complex problem solving tasks. *Technol. Knowl. Learn.* 19, 127–146. doi: 10.1007/s10758-014-9222-8

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Dörner and Funke. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# The Role of Motivation in Complex Problem Solving

C. Dominik Güss<sup>1\*</sup>, Madison Lee Burger<sup>1</sup> and Dietrich Dörner<sup>2</sup>

<sup>1</sup> Department of Psychology, University of North Florida, Jacksonville, FL, United States, <sup>2</sup> Trimberg Research Academy, University of Bamberg, Bamberg, Germany

**Keywords:** complex problem solving, dynamic decision making, simulation, motivation, PSI-theory, Maslow's hierarchy of needs, achievement motivation

## THE ROLE OF MOTIVATION IN COMPLEX PROBLEM SOLVING

Previous research on Complex Problem Solving (CPS) has primarily focused on cognitive factors as outlined below. The current paper discusses the role of motivation during CPS and argues that motivation, emotion, and cognition interact and cannot be studied in an isolated manner. Motivation is the process that determines the energization and direction of behavior (Heckhausen, 1991).

Three motivation theories and their relation to CPS are examined: McClelland's achievement motivation, Maslow's hierarchy of needs, and Dörner's needs as outlined in PSI-theory. We chose these three theories for several reasons. First, space forces us to be selective. Second, the three theories are among the most prominent motivational theories. Finally, they are need theories postulating several motivations and not just one. A thinking-aloud protocol is provided to illustrate the role of motivational and cognitive dynamics in CPS.

Problems are part of all the domains of human life. The field of CPS investigates problems that are complex, dynamic, and non-transparent (Dörner, 1996). Complex problems consist of many interactively interrelated variables. Dynamic ones change and develop further over time, regardless of whether the involved people take action. And non-transparent problems have many aspects of the problem situation that are unclear or unknown to the involved people.

CPS researchers focus exactly on such kinds of problems. Under a narrow perspective, CPS can be defined as thinking that aims to overcome barriers and to reach goals in situations that are complex, dynamic, and non-transparent (Frensch and Funke, 1995). Indeed, past research has shown the influential role of task properties (Berry and Broadbent, 1984; Funke, 1985) and of cognitive factors on CPS strategies and performance, such as intelligence (e.g., Süß, 2001; Stadler et al., 2015), domain-specific knowledge (e.g., Wenke et al., 2005), cognitive biases and errors (e.g., Dörner, 1996; Güss et al., 2015), or self-reflection (e.g., Donovan et al., 2015).

Under a broader perspective, CPS can be defined as the study of cognitive, emotional, motivational, and social processes when people are confronted with such complex, dynamic and non-transparent problem situations (Schoppek and Putz-Osterloh, 2003; Dörner and Güss, 2011, 2013; Funke, 2012). The assumption here is that focusing solely on cognitive processes reveals an incomplete picture or an inaccurate one.

To study CPS, researchers have often used computer-simulated problem scenarios also called microworlds or virtual environments or strategy games. In these situations, participants are confronted with a complex problem simulated on the computer from which they gather information, and identify solutions. These decisions are then implemented into the system and result in changes to the problem situation.

## PREVIOUS RESEARCH ON MOTIVATION AND CPS

The idea to study the interaction of motivation, emotion, and cognition is not new (Simon, 1967). However, in practice this has been rarely examined in the field of CPS. One study assessed the need

### OPEN ACCESS

#### Edited by:

Joachim Funke,  
Heidelberg University, Germany

#### Reviewed by:

Christine Blech,  
FernUniversität Hagen, Germany  
Ricarda Steinmayr,  
Technische Universität Dortmund,  
Germany

#### \*Correspondence:

C. Dominik Güss  
dguss@unf.edu

#### Specialty section:

This article was submitted to  
Cognitive Science,  
a section of the journal  
Frontiers in Psychology

**Received:** 16 March 2017

**Accepted:** 09 May 2017

**Published:** 23 May 2017

#### Citation:

Güss CD, Burger ML and Dörner D  
(2017) The Role of Motivation in  
Complex Problem Solving.  
Front. Psychol. 8:851.  
doi: 10.3389/fpsyg.2017.00851



for cognition (i.e., the tendency to engage in thinking and reflecting) and showed how high need of cognition was related to broader information collection and better performance in a management simulation (Nair and Ramnarayan, 2000).

Vollmeyer and Rheinberg (1999, 2000) explored in two studies the role of motivational factors in CPS. They assessed mastery confidence (similar to self-efficacy), incompetence fear, interest, and challenge as motivational factors. Their results demonstrated that mastery confidence and incompetence fear were good predictors for learning and for knowledge acquisition.

## CPS ASSESSMENT

Before we describe three theories of motivation and how they might be related and applicable to CPS, we will briefly describe the WINFIRE computer simulation (Gerdes et al., 1993; Schaub, 2009) and provide a part of a thinking-aloud protocol of one participant while working on WINFIRE. WINFIRE is the simulation of small cities surrounded by forests. Participants take the role of fire-fighting commanders who try to protect cities and forests from approaching fires. Participants can give a series of commands to several fire trucks and helicopters. In WINFIRE quick decisions and multitasking are required in order to avoid fires spreading. In one study, participants were also instructed to think aloud, i.e., to say aloud everything that went through their minds while working on WINFIRE. These thinking-aloud protocols, also called verbal protocols, were audiotaped and transcribed in five countries and compared (see Güss et al., 2010).

The following is a verbatim WINFIRE thinking-aloud protocol of a US participant (Güss et al., 2010):

*Ok, I don't see any fires yet. I'm trying to figure out how the helicopters pick up the water from the ponds. I put helicopters on patrol mode. Not really sure what that does. It doesn't seem to be moving. Oh, there it goes, it's moving... I guess you have to wait till there's a fire showing... Ok, fire just started in the middle, so I have to get some people to extinguish it. Ok, now I have another fire going here. I'm in trouble here. Ok. Ok, when I click extinguish, it don't seem to respond. Guess I'm not clear how to get trucks right to the fire. Ok, one fire has been extinguished, but a new one started in the same area. I'm getting more trucks out there trying to figure out, how to get helicopters to the pond. I still haven't figured that out, because they have to pick up the water. Ok, got a pretty good fire going here, so I'm going to put all the trucks on action, ok, water thing is making me mad. Ok. I'm not sure how it goes? Ok, the forest is burning up now—extinguish! Ok, ok, I'm in big trouble here...*

## PSYCHOLOGICAL THEORIES OF MOTIVATION AND THEIR APPLICATION TO CPS

### McClelland's Human Motivation Theory

In his Human Motivation theory, McClelland distinguishes three needs (power, affiliation, and achievement) and argues that human motivation is a response to changes in affective states. A specific situation will cause a change in the affective state through the non-specific response of the autonomic nervous system.

This response will motivate a person toward a goal to reach a different affective state (McClelland et al., 1953). An affective state may either be positive or negative, determining the direction of motivated behavior as either approach oriented, i.e., to maintain the state, or avoidance oriented, i.e., to avoid or discontinue the state (McClelland et al., 1953).

Motivation intensity varies among individuals based on perception of the stimulus and the adaptive abilities of the individual. Hence, when a discrepancy exists between expectation and perception, then a person will be motivated to eliminate this discrepancy (McClelland et al., 1953). In the statement from the thinking-aloud protocol we can infer the participant's achievement motivation, "Guess I'm not clear how to get trucks right to the fire. Ok, one fire has been extinguished, but a new one started in the same area." The participant at first begins to give up and reduce effort, but then achieves a step toward the goal. This achievement causes the reevaluation of the discrepancy between ability and the goal as not too large to overcome. This realization motivates the participant to continue working through the scenario. Whereas, the need for achievement seems to guide CPS, the needs for power and affiliation cannot be observed in the current thinking-aloud protocol.

Based on the previous discussion we can derive the following predictions:

*Prediction 1:* Approach-orientation will lead to greater engagement in CPS compared to avoidance-orientation.

*Prediction 2:* Based on an individual's experience either power, affiliation, or achievement will become dominant and guide the strategic approach in CPS.

### Maslow's Hierarchy of Needs

Maslow's Hierarchy of Needs (Maslow, 1943, 1954) suggests that everyone has five basic needs that act as motivating forces in a person's life. Maslow's hierarchy takes the form of a pyramid in which needs lower in the pyramid are primary motivators. They have to be met before higher needs can become motivating forces. At the bottom of the pyramid are the most basic needs beginning with physiological needs, such as hunger, and followed by safety needs. Then follow the psychological needs of belongingness and love, and then esteem. Once these four groups of needs have been met, a person may reach the self-fulfillment stage of self-actualization at which time a person can be motivated to achieve ones full potential (Maslow, 1943).

The first four groups of needs are external motivators because they motivate through both deficiency and fulfillment. In essence, a person fulfills a need which then releases the next unsatisfied need to be the dominant motivator (Maslow, 1943, 1954). The safety need is often understood as seeking shelter, but Maslow also understands safety also as wanting "a predictable, orderly world" (Maslow, 1943, p. 377), "an organized world rather than an unorganized or unstructured one" (Maslow, 1943, p. 377). Safety refers to the "common preference for familiar rather than unfamiliar things" (Maslow, 1943, p. 379).

In this sense the safety need becomes active when the person does not understand what is happening in the microworld, as

the following passage of the thinking-aloud protocol illustrates. *“I put helicopters on patrol mode. Not really sure what that does. It doesn’t seem to be moving.”* The safety need is demonstrated in the person’s desire for organization, since unknown and unexpected events are seen as threats to safety.

The esteem need as a motivator becomes evident through the statement, *“Guess I’m not clear how to get trucks right to the fire.”* The participant becomes aware of his inability to control the situation which affects his self-esteem. The esteem need is never fulfilled in the described situation and remains the primary motivator. The following statements show how affected the participant’s esteem need is by the inability to control the burning fires. *“Ok. I’m not sure how it goes? Ok, the forest is burning up now—extinguish! Ok, ok, I’m in big trouble.”*

**Prediction 3:** A strong safety need will be related to elaborate and detailed information collection in CPS compared to low safety need.

**Prediction 4:** People with high esteem needs will be affected more by difficulties in CPS and engage more often in behaviors to protect their esteem compared to people with low esteem needs.

## Dörner’s Theory of Motivation as Part of PSI-Theory

PSI-theory described the interaction of cognitive, emotional, and motivational processes (Dörner, 2003; Dörner and Güss, 2011). Only a small part of the theory is examined here. Briefly, the theory encompasses five basic human needs: the existential needs (thirst, hunger, and pain avoidance), the sexuality need, and the social need for affiliation (group binding), the need for certainty (predictability), and the need for competence (mastery). If the environment is unpredictable, the certainty need becomes active. If we are not able to cope with problems, the competence need becomes active. The need for competence also becomes active when any other need becomes activated. With an increase in needs, the arousal increases.

The first three needs cannot be observed or inferred from the thinking-aloud protocol provided. Statements like, *“I’m trying to figure out how the helicopters pick up the water from the ponds.”* and *“Guess I’m not clear how to get trucks right to the fire,”* demonstrate the needs for certainty and competence, i.e., to make the environment predictable and controllable.

The following statements reflect the participant’s need for competence, i.e., the inefficacy or incapability of coping with problems. *“I’m in trouble here... ok, water thing is making me mad.”* Not being able to extinguish the fires that are approaching cities and are destroying forests is experienced as anger. The arousal rises as the resolution level of thinking decreases. So, the participant does not think about different options in an elaborate manner. Yet, the participant becomes aware of his failure. The competence need then causes the participant to search for possible solutions, *“I still haven’t figured that out because they have to pick up the water...”* The need for competence is satisfied when the problem solver is able to change either the environment or one’s views of the environment.

**Prediction 5:** A strong certainty need is positively related to a strong competence need.

**Prediction 6:** High need for certainty paired with high need for competence can lead to safeguarding behavior, i.e., background monitoring.

**Prediction 7:** An increase in the competence and uncertainty needs leads to increased arousal and a lower resolution level of thinking. CPS becomes one-dimensional and possible long-term and side-effects are not considered adequately.

## Summary and Evaluation

We have briefly discussed three motivation theories and their relation to CPS referring to one thinking-aloud protocol: McClelland’s achievement motivation, Maslow’s hierarchy of needs, and Dörner’s needs as outlined in PSI-theory.

A Comparison of Three Need Theories in the Context of CPS.

	McClelland’s achievement motivation	Maslow’s hierarchy of need	Dörner’s theory of motivation as part of PSI-theory
Scope/ Breadth	+	++	++
Applicability to CPS	+	+	++
Adaptability of needs	—	—	+
Incorporation of emotion	++	—	++
Individual differences	++	—	++

*Evaluation criteria: very small/very low —, small/low —, much/high +, very much/very high ++.*

Comparing the scope of the three theories and referring to the scope and different needs covered in the three theories, McClelland’s theory describes three needs (power, affiliation, and achievement), Maslow’s theory describes five groups of needs (physiological, safety, love and belonging, esteem, self-actualization), and Dörner’s theory describes five different needs (existential, sexuality, affiliation, certainty, and competence).

All three theories can be applied to CPS. McClelland’s need for achievement, Maslow’s needs for esteem and safety, and Dörner’s needs for certainty and competence could be inferred from the thinking-aloud passage. The need for affiliation which is a part of each of the three theories could play an important role when groups solve complex problems.

The existential needs and the need for affiliation outlined in PSI-theory can also be found in Maslow’s hierarchy of needs. These two theories differ in the adaptability of the needs. However, Maslow’s esteem needs are only activated as the primary motivator as the physiological needs, belongingness, and love needs are met. The needs are more fluidly described as motivators in PSI-theory. One need becomes the dominant

motive according to the expectancy–value principle. Expectancy stands for the estimated likelihood of success. The value of a motive stands for the strength of the need. According to McClelland's theory, the role of three motivations develops through life experience in a specific culture; and often times, one of the three becomes the main driving force for a person, almost like a personality trait. In that sense, there is not much flexibility.

Motivation and emotion are closely related as became partially clear in the discussion of McClelland's theory. Emotions are discussed in detail in PSI-theory, but space does not allow us to discuss those in detail here (see Dörner, 2003). Emotions are not described in detail in Maslow's Hierarchy of Needs.

Individual differences in motivation and needs are discussed in two of the three theories. According to McClelland, a person develops an individual achievement motive by learning one's own abilities from past achievements and failures. Based on different learning histories, different persons will have a different dominant motivation guiding behavior in a given situation. Learning history also influences the competence need in PSI-theory. Additionally PSI-theory assumes individual differences that are simulated through different individual motivational parameters in the theory. The certainty need, for example, becomes active when there is a deviation from a given set point. Individual differences are related to different set points and how sensitive the deviations are (e.g., deviation starts quickly vs. deviation starts slowly).

## REFERENCES

- Berry, D. C., and Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *Q. J. Exp. Psychol.* 36A, 209–231. doi: 10.1080/14640748408402156
- Donovan, S., Güss, C. D., and Naslund, D. (2015). Improving dynamic decision making through training and self-reflection. *Judgm. Decis. Making* 10, 284–295.
- Dörner, D. (1996). *The logic of failure. Recognizing and Avoiding Error in Complex Situations*. New York, NY: Basic Books.
- Dörner, D. (2003). *Bauplan für eine Seele [Blueprint for a soul]*. Reinbek: Rowohlt.
- Dörner, D., and Güss, C. D. (2011). A psychological analysis of Adolf Hitler's decision making as Commander in Chief: *summa confidentia et nimius metus*. *Rev. Gen. Psychol.* 15, 37–49. doi: 10.1037/a0022375
- Dörner, D., and Güss, C. D. (2013). PSI: a computational architecture of cognition, motivation, and emotion. *Rev. Gen. Psychol.* 17, 297–317. doi: 10.1037/a0032947
- Eccles, J. S., and Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annu. Rev. Psychol.* 53, 109–132. doi: 10.1146/annurev psych.53.100901.135153
- Frensch, P., and Funke, J. (eds.) (1995). *Complex Problem Solving: The European Perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
- Funke, J. (1985). Steuerung dynamischer Systeme durch Aufbau und Anwendung subjektiver Kausalmodelle [Control of dynamic systems via Construction and Application of subjective causal models]. *Zeitschrift für Psychol.* 193, 443–465.
- Funke, J. (2012). "Complex problem solving," in *Encyclopedia of the Sciences of Learning*, ed N. M. Seel (Heidelberg: Springer), 682–685.
- Gerdes, J., Dörner, D., and Pfeiffer, E. (1993). *Interaktive Computersimulation "Winfire."* [The Interactive Computer Simulation "Winfire"]. Otto-Friedrich-Universität Bamberg: Lehrstuhl Psychologie, II.
- Güss, C. D., Tuason, M. T., and Gerhard, C. (2010). Cross-national comparisons of complex problem-solving strategies in two microworlds. *Cogn. Sci.* 34, 489–520. doi: 10.1111/j.1551-6709.2009.01087.x
- Güss, C. D., Tuason, M. T., and Orduña, L. V. (2015). Strategies, tactics, and errors in dynamic decision making. *J. Dyn. Decis. Making* 1, 1–14. doi: 10.11588/jddm.2015.1.13131
- Heckhausen, H. (1991). *Motivation and Action*. New York, NY: Springer.
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Maslow, A. H. (1954). *Motivation and Personality*. New York, NY: Harper and Row.
- McClelland, D. C., Atkinson, J. W., Clark, R. A., and Lowell, E. L. (1953). *The Achievement Motive*. East Norwalk, CT: Appleton-Century-Crofts.
- Nair, K. U., and Ramnarayan, S. (2000). Individual differences in need for cognition and complex problem solving. *J. Res. Pers.* 34, 305–328. doi: 10.1006/jrpe.1999.2274
- Schaub, H. (2009). *Fire Simulation*. Ottobrunn: IABG.
- Schoppek, W., and Putz-Osterloh, W. (2003). Individuelle Unterschiede und die Bearbeitung komplexer Probleme. *Zeitschrift für Differentielle und Diagnostische Psychol.* 24, 163–173. doi: 10.1024/0170-1789.24.3.163
- Simon, H. A. (1967). Motivational and emotional controls of cognition. *Psychol. Rev.* 74, 29–39. doi: 10.1037/h0024127
- Süß, H.-M. (2001). "Die Rolle von Intelligenz und Wissen für erfolgreiches Handeln in komplexen Problemsituationen [The role of intelligence and knowledge for successful performance in complex problem solving]," in *Komplexität und Kompetenz: Ausgewählte Fragen der Kompetenzforschung*, ed G. Franke (Bielefeld: Bertelsmann), 249–275.
- Stadler, M., Becker, N., Gödker, M., Leutner, D., and Greiff, S. (2015). Complex problem solving and intelligence: a meta-analysis. *Intelligence* 53, 92–101. doi: 10.1016/j.intell.2015.09.005

## CONCLUSION

The thinking-aloud example from the WINFIRE microworld described earlier demonstrates that a person's CPS process is influenced by the person's needs. We have focused in our discussion on motivational processes that are considered in the framework of need theories. Beyond that, other motivational theories exist that focus on the importance of motivation for learning and achievement (e.g., expectancy, reasons for engagement, see Eccles and Wigfield, 2002). Thus, the applicability of these theories to CPS could be explored in future studies as well.

We discussed the three motivational theories of McClelland's Achievement Motivation, Maslow's Hierarchy of Need, and Dörner's Theory of Motivation as part of PSI-Theory. Although, the theories differ our discussion has shown that the three theories can be applied to CPS. Problem solving is a motivated process and determined by human motivations and needs.

## AUTHOR CONTRIBUTIONS

The first author CG conceptualized the manuscript, selected the thinking-aloud passage, the second author MB primarily summarized McClelland's and Maslow's theories. All authors contributed to writing up the manuscript.

- Vollmeyer, R., and Rheinberg, F. (1999). Motivation and metacognition when learning a complex system. *Eur. J. Psychol. Educ.* 14, 541–554. doi: 10.1007/BF03172978
- Vollmeyer, R., and Rheinberg, F. (2000). Does motivation affect performance via persistence? *Learn. Instruct.* 10, 293–309. doi: 10.1016/S0959-4752(99)00031-6
- Wenke, D., Frensch, P. A., and Funke, J. (2005). “Complex problem solving and intelligence: empirical relation and causal direction,” in *Cognition and Intelligence: Identifying the Mechanisms of the Mind*, eds R. J. Sternberg and J. E. Pretz (New York, NY: Cambridge University Press), 160–187.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Güss, Burger and Dörner. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Approaches to Cognitive Modeling in Dynamic Systems Control

Daniel V. Holt<sup>1\*</sup> and Magda Osman<sup>2</sup>

<sup>1</sup> Department of Psychology, Heidelberg University, Heidelberg, Germany, <sup>2</sup> Department of Biological and Experimental Psychology, School of Biological and Chemical Sciences, Queen Mary University of London, London, United Kingdom

## OPEN ACCESS

### Edited by:

Rick Thomas,  
Georgia Institute of Technology,  
United States

### Reviewed by:

Varun Dutt,  
Indian Institute of Technology Mandi,  
India  
Shenghua Luan,  
Max Planck Institute for Human  
Development, Germany

### \*Correspondence:

Daniel V. Holt  
daniel.holt@psychologie.  
uni-heidelberg.de

### Specialty section:

This article was submitted to  
Cognitive Science,  
a section of the journal  
Frontiers in Psychology

**Received:** 15 March 2017

**Accepted:** 06 November 2017

**Published:** 29 November 2017

### Citation:

Holt DV and Osman M (2017)  
Approaches to Cognitive Modeling  
in Dynamic Systems Control.  
Front. Psychol. 8:2032.  
doi: 10.3389/fpsyg.2017.02032

Much of human decision making occurs in dynamic situations where decision makers have to control a number of interrelated elements (dynamic systems control). Although in recent years progress has been made toward assessing individual differences in control performance, the cognitive processes underlying exploration and control of dynamic systems are not yet well understood. In this perspectives article we examine the contribution of different approaches to modeling cognition in dynamic systems control, including instance-based learning, heuristic models, complex knowledge-based models and models of causal learning. We conclude that each approach has particular strengths in modeling certain aspects of cognition in dynamic systems control. In particular, Bayesian models of causal learning and hybrid models combining heuristic strategies with reinforcement learning appear to be promising avenues for further work in this field.

**Keywords:** dynamic decision making, complex problem solving, cognitive modeling, instance-based learning, heuristics, causal learning

## INTRODUCTION

Handling dynamic systems is a common requirement in everyday life, ranging from operating a novel technical device, to managing a company or understanding the dynamics of social relationships. Considerable progress has been made toward assessing dynamic systems control (DSC) as a cognitive skill, particularly in educational contexts (OECD, 2014; Schoppek and Fischer, 2015). However, the development of cognitive theories which describe and explain the mental processes underlying this skill seems to lag behind. In this perspectives article we therefore examine what computational models of cognition can contribute toward an improved understanding of different cognitive processes involved in DSC. For this purpose, we briefly review several approaches to cognitive modeling in DSC, summarize their relative strengths and weaknesses, and conclude with what we perceive as promising routes for future research.

Dynamic systems control can be defined as a form of dynamic decision making that requires (1) a series of interrelated decisions (2) in interaction with a dynamic system inducing states of subjective uncertainty (3) with the aim of attaining (and maintaining) a goal state and/or to explore the system and possible courses of action (also see Edwards, 1962; Osman, 2010). Subjective uncertainty may be caused by random fluctuations of the system but also by limited knowledge of the system's structure and its dynamics (Osman, 2010). In cognitive research DSC is typically investigated using computer-simulated *microworlds*, which are the focus of the present paper (but see Klein et al., 1993, for a different approach). Microworlds emulate cognitively relevant features of DSC situations (e.g., limited information, delayed feedback, time pressure) framed in a semantically plausible setting such as managing a company or fighting a forest fire (Brehmer and Dörner, 1993; Gonzalez et al., 2005).

In this article, the terms *cognitive model* or *computational model (of cognition)* refer to any mechanistic account of cognitive processes that is sufficiently specified to allow a computer-based implementation yielding quantitative predictions of behavior, cognitive processing steps, or neural activity (Lewandowsky and Farrell, 2011). Computational models enforce conceptual completeness, as all functionally relevant properties of a theory have to be made explicit for its computational implementation. Furthermore, computational models generate precise predictions of data patterns that can be empirically tested, which most verbally expressed theories are unable to do with the same level of precision. Models implemented as computer programs can also be used for simulation-based exploration to investigate how varying parameter settings, assumptions about cognitive processes, or simulated task demands affect model behavior. For a comprehensive treatment of computational modeling in cognition – including its challenges and problems – please refer to Lewandowsky and Farrell (2011).

We will now consider the contribution of different approaches to modeling cognition in DSC with selected examples. Our brief review begins with comparatively simple and knowledge-lean modeling paradigms, moving toward approaches involving more complex strategies and knowledge structures (see **Table 1** for an overview).

## INSTANCE-BASED LEARNING

Instance-based learning (IBL) models have arguably been among the most successful approaches to cognitive modeling in DSC. They are based on a simple principle analogous to reinforcement learning: control actions that lead to a successful outcome become reinforced in memory and will therefore more likely be remembered and enacted when a similar situation is encountered in the future (Gonzalez et al., 2003). The application of this learning principle in DSC can be traced back to Berry and Broadbent's (1984) semi work on knowledge acquisition in dynamic systems. In their deceptively simple task, the production of a simulated sugar factory had to be kept in a target range by adjusting the number of workers. Surprisingly, most participants were unable to verbally describe the system's structure despite being able to control it above chance level. As an explanation, Broadbent et al. (1986) suggested that people might store instances of situation-action combinations they have experienced in memory (e.g., hiring X workers when the current number of workers is Y and the current production is Z). Subsequent decision making in turn is based on retrieving instance memories similar to the situation encountered. Those instances repeatedly associated with successful outcomes (e.g., reaching the target production) become reinforced in memory and gradually start to dominate behavior, although no verbalizable representation of the system's structure has been formed.

A computational IBL model of the sugar factory task has been implemented by Taatgen and Wallach (2002; also see Dienes and Fahey, 1995) using the ACT-R cognitive architecture (Anderson et al., 2004). Each instance was modeled as a unit in declarative memory encoding current state, action

and outcome. On encountering a given system state, the model searches for instance memories that are similar to the current state and have led to the target outcome in the past, taking into account how often the instance memory has been retrieved before. The model requires only two production rules and closely fits human behavior. A model of the sugar factory relying on a similar associative learning mechanism was implemented by Gibson et al. (1997) using an artificial neural network. This illustrates that the basic learning principle is independent of any specific modeling architecture. IBL has also been applied to modeling more complex tasks such as controlling an array of pumps in a simulated water purification plant where the system state changes in real time (Gonzalez et al., 2003). The model included a blending mechanism to interpolate information across related instances and relied on a simple decision heuristic as fallback when instance memories were insufficient. A generic implementation of the IBL framework has been made available by Dutt and Gonzalez (2012) to make IBL modeling accessible to non-expert modelers.

An approach similar to IBL was used by Glass and Osman (2017; also Osman et al., 2015) to model learning in a simple dynamic system with continuous input and output variables. Instead of relying on a cognitive architecture, the authors adapted a general-purpose reinforcement learning algorithm to this task (Sutton and Barto, 1998). The model updated the reinforcement history of input variables after each trial depending on how much the last action reduced goal distance. This results in model behavior broadly similar to IBL. Glass and Osman (2017) specifically focused on modeling group differences between young and old adults in terms of exploration vs. exploitation behavior. They mapped this behavioral preference on the noise parameter affecting the choice of input values. Reinforcement learning has also been used to model conflicts between short- and long-term goals and how unreliable information affects learning in dynamic control (Gureckis and Love, 2009).

In sum, IBL and reinforcement learning models have been successfully used to explain different aspects of exploration and control in DSC. The basic mechanism is simple, cognitively plausible and requires only few task-specific assumptions. However, IBL models critically depend on the availability of immediate outcome feedback and the frequent repetition of similar decision situations to facilitate learning, which limits the type of task they can be applied to (see **Table 1**). Furthermore, IBL models cannot easily explain how people acquire explicit knowledge of the causal structure of a system, which is a central element of some DSC tasks (e.g., Kluge, 2008; Wüstenberg et al., 2012).

## HEURISTIC MODELS

Heuristics-based approaches to DSC assume that people rely on simple rule-of-thumb-type decision strategies for controlling dynamic systems. These strategies do not guarantee an optimal result, but allow to achieve reasonable outcomes across a range of conditions with limited cognitive effort (Brehmer, 2005; Shah

and Oppenheimer, 2008). Characteristically, heuristic strategies do neither involve complex reasoning nor a detailed mental representation of the problem structure. Empirical research has shown that heuristics can explain adaptive behavior in many decision making situations as well as common errors and biases (Gilovich et al., 2002; Gigerenzer and Brighton, 2009). Furthermore, due to their simplicity heuristics are relatively easy to implement as cognitive models (e.g., Marewski and Mehlhorn, 2011).

The use of heuristics has also been proposed as an explanation for decision making behavior in DSC (e.g., Brehmer and Elg, 2005; Cronin et al., 2009). One of the best known examples of a computational heuristic model in DSC is Stermann's (1989) model of decision making in a supply chain management task. The model is based on the anchoring-and-adjustment heuristic (Tversky and Kahneman, 1974), which involves substituting an unknown quantity (supplies ordered) with a related known quantity (sales forecast), adjusting for further influences (current stock level). Data simulated using this heuristic closely match human behavior and reproduce the characteristic oscillation between over- and undersupply that arises from ignoring system delays (Stermann, 1989). The cold store temperature regulation task (Reichert and Dörner, 1988) poses a similar challenge to participants, as the system responds with delay to changed inputs. Reichert and Dörner's (1988) model captures how participants gradually learn to control the system by applying incremental changes to a proportional-control heuristic after unsuccessful control attempts. A conceptually related adaptive heuristic strategy is directional learning: if increasing an input improves the outcome then continue to increase it, otherwise decrease it. Computational models of directional learning have for example been used to model behavior in dynamic economic

games, such as the multiple-round prisoner's dilemma or the ultimatum game (Selten and Stoecker, 1986; Grosskopf, 2003).

Heuristic models can also be combined with reinforcement learning to simulate how people learn to choose among competing heuristic strategies. The probability of selecting a strategy depends on the outcomes that this strategy has produced in the past (Erev and Barron, 2005). For example, Gonzalez et al. (2009) used this approach to model response times in a dynamic radar detection and decision making task. The model fitted human data about as well as an alternative IBL model, although it transferred less well to changed task conditions. In contrast, Fum and Stocco (2003) found that a strategy-based learning model of the sugar factory task performed better under changed conditions than the corresponding IBL model of Taatgen and Wallach (2002). It appears that the transfer across situations depends on the details of the task, the type of training, and the strategies implemented. Strategy-based learning can also be applied to highly complex tasks such as fighting a simulated forest fire (De Obeso Orendain and Wood, 2012). In this model four high-level heuristic strategies competed (e.g., dropping water on the fire or creating a barrier to contain the fire), which were modeled in great detail. The model successfully reproduced how varying the conditions in the training phase affected preferences for particular strategies in later transfer.

In sum, there are several good examples of heuristics-based and hybrid models that address pertinent theoretical questions in DSC (e.g., handling delays, transfer of learning). Although incorporating more task-specific knowledge than pure learning models, models of this type typically still have a relatively simple basic structure. When heuristic models are extended with more complex strategies and abstract knowledge structures

**TABLE 1 |** Basic approaches to cognitive modeling in dynamic systems control (DSC).

Approach	Requirements	Strengths	Limitations	Examples
<i>Instance-based learning:</i> Acquisition of situation-response associations guided by outcome feedback.	Frequent exposure to similar states of the DSC task. Prior knowledge can be minimal.	Simple formalism Universal applicability across different domains High neural and cognitive plausibility	No representation of causal knowledge Requires direct outcome feedback	Taatgen and Wallach, 2002; Gonzalez et al., 2003
<i>Heuristic models:</i> Simple "rules of thumb" with low cognitive requirements.	Situations for which control heuristics based on prior knowledge are available.	Structurally simple models Role of heuristics in decision making well supported Can be extended with reinforcement learning	No representation of causal knowledge Need to establish suitable heuristics for each domain	Reichert and Dörner, 1988; Stermann, 1989
<i>Complex knowledge-based models:</i> Complex cognitive strategies and abstract mental representations.	Typically requires considerable task-specific prior knowledge.	Modeling of complex knowledge structures and reasoning strategies	Can be very complex Often highly task-specific, limited transfer	Schunn and Anderson, 1998; Schoppek, 2002
<i>Causal learning:</i> Bayesian induction of causal relations from observing system behavior.	Task sufficiently simple to allow causal attribution. Prior knowledge can be minimal.	Comprehensive formalism for representing causal knowledge, uncertainty and knowledge updating	Have not yet been directly applied to system control tasks	Steyvers et al., 2003; Meder et al., 2010

they gradually transform into complex knowledge-based models, which we will consider next.

## COMPLEX KNOWLEDGE-BASED MODELS

Complex knowledge-based models involve the creation and transformation of abstract knowledge structures combined with complex cognitive strategies. This corresponds to the notion of mental models guiding reasoning and actions in DSC (e.g., Brehmer, 1992). Models of this type are often implemented as production systems, i.e., sets of if-then-rules that act on knowledge objects or initiate external behaviors (Newell and Simon, 1972). This flexible mechanism allows to express a wide range of strategies up to expert performance at the expense of resulting in very complex and task-specific models in many cases.

Anzai (1984) presented one of the first models of this type in the context of navigating a simulated large ship which responded with considerable delay. Based on the analysis of verbal protocols, Anzai (1984) designed a production system model that qualitatively captured the acquisition of control knowledge in novices and experts. Later studies showed how production system approaches can be extended to model performance even in complex real-time decision making tasks such as that of a radar operator or flying a commercial jet plane (Schoelles and Gray, 2000; Schoppek and Boehm-Davis, 2004). In modeling real-time control tasks the emphasis usually lies on modeling the time course of control performance and typical errors.

A different class of DSC task that requires scientific reasoning to explore the causal structure of unknown systems has been particularly relevant in the recent educationally oriented wave of DSC research (Herde et al., 2016). Schoppek (2002) proposed a fine-grained cognitive model that was able to systematically explore and control a small dynamic system based on linear equations. The model incorporated an explicit mental representation of the system's structure and mental calculation steps to derive input values. This strategy is sufficiently general to control any simple dynamic system based on linear equations (Funke, 2001). The model was able to simulate the effects of different degrees of system knowledge and strategic sophistication and compared favorably to results from human data (Schoppek, 2002). Similarly, Schunn and Anderson (1998) applied a production system approach to model the task of designing scientific experiments (given a restricted set of design options) and drawing conclusions about causal relations from the simulated results. This model was able to successfully capture performance differences between experts and novices by modeling their respective domain knowledge and exploration strategies.

An apparent advantage of complex knowledge-based models is their ability to explain how causal system knowledge combined with reasoning strategies informs the actions that people take. It seems difficult to imagine how some forms of DSC could be explained without recourse to reasoning and abstract knowledge representation, for example, extrapolating system behavior in new situations or hypothesis testing and rule-deduction in

discovery learning. However, models of this type are often neither simple nor elegant and require the inclusion of considerable task-specific knowledge (Taatgen and Anderson, 2010).

## BAYESIAN CAUSAL LEARNING

Another modeling approach with a focus on structural knowledge is the use of Bayesian networks to model causal learning (Meder et al., 2010; Osman, 2017). Bayesian networks represent a formalism to express the strength of belief in causal hypotheses and provide a principled mechanism based on Bayesian inference for updating beliefs as new evidence becomes available (Holoak and Cheng, 2011). For instance, Steyvers et al. (2003) used Bayesian networks to model human causal learning either by passive observation of a causal system or through direct interaction with it. This addresses a central aspect of DSC, system exploration and the formation of structural knowledge, although the approach has yet to be applied to DSC tasks requiring goal-directed control.

From a DSC perspective, the strength of Bayesian models of causal learning lies in the nuanced representation of causal structures and probabilistic dependencies combined with a mechanism for updating this knowledge from experience. This makes them a strong contender for explaining structural knowledge acquisition in DSC tasks with an exploration focus (e.g., Kluge, 2008; Wüstenberg et al., 2012). A Bayesian approach provides a formal account of the causal environment from which it is possible to deduce a suitable course of action, given the state of knowledge (including the level of uncertainty) a person has of the world at that time (Osman, 2010, 2017).

## SYNTHESIS AND CONCLUDING REMARKS

In pursuit of answering the question what cognitive modeling can contribute to DSC research we have considered several approaches (see **Table 1**). In terms of knowledge-lean modeling approaches, IBL strikes a good balance between simplicity, cognitive plausibility and explanatory power for a range of DSC tasks. On the downside, IBL has strict task requirements (e.g., availability of feedback, repeated decisions) and cannot easily explain the acquisition of causal knowledge or extrapolation to unfamiliar conditions. Heuristic models have no universal task requirements and can be combined with learning mechanisms to achieve a similar adaptivity as IBL models. However, since effective heuristics rely on exploiting the structure of the environment, finding suitable candidate heuristics for a given task can be a considerable challenge and any specific heuristics-based model is only applicable to a particular niche (Marewski and Schooler, 2011).

Complex knowledge-based models are probably the most domain-specific type of model. They require strong assumptions about knowledge structures and cognitive procedures used by decision makers. If this information is available, it is possible to



model skilled expert performance, elaborate reasoning strategies, and the acquisition of explicit structural knowledge, e.g., through active hypothesis testing. With respect to modeling causal knowledge, Bayesian models of causal learning provide an interesting alternative.

They offer an integrated account for representing and updating causal knowledge in a coherent framework, including the representation of epistemic uncertainty. However, these models have so far not been directly applied to model control in microworld DSC tasks.

As our discussion shows, each modeling approach has its particular strengths and weaknesses, which render it suitable for particular modeling tasks. For researchers it therefore seems important to select a modeling approach that matches the research question and that suits the task to be modeled. For example, modeling the process of acquiring explicit causal knowledge in simple dynamic systems through hypothesis testing (e.g., Wüstenberg et al., 2012) naturally maps on complex knowledge-based models or Bayesian models of causal learning but is likely to run into difficulties when approached with an IBL framework.

In general, we think that the rapidly advancing theories in causal learning and hybrid models combining heuristic strategies with reinforcement learning offer considerable untapped potential for cognitive modeling in DSC. Causal learning directly addresses the core issue of DSC tasks focusing on causal exploration (e.g., Steyvers et al., 2003). However, in order to simulate full task performance models of this type would need to be extended by including interaction with the task. Hybrid models in turn may be most suitable to model behavior in complex decision making tasks (e.g., Danner et al., 2011), where (heuristically guided) information reduction and gradual strategy adaptation are central for task performance.

We furthermore propose that computational models based on cognitively plausible process assumptions (e.g., reinforcement

learning, use of simple heuristics, Bayesian knowledge updating) could be used as a yardstick for evaluating human performance in DSC (see Brehmer, 2005). This stands in contrast to using mathematical optimization or optimal rational strategies as a benchmark for performance (e.g., Sager et al., 2011). Defining rationally optimal strategies in DSC does have its place, for example when designing decision support systems. However, from a behavioral perspective the question of “what is *maximally* possible” is often less relevant than “what is *humanly* possible,” given the realities of incomplete information and limited cognitive capacity (Klein, 2002).

In conclusion, computational models of cognition appear to offer a promising path for advancing research and theory development in DSC. Computational approaches have successfully been used to model a range of cognitive phenomena in different domains of DSC. Promising starting points for further developments include, for example, recent advances in causal learning and hybrid models which combine simple heuristics with reinforcement learning mechanisms. Computational modeling of cognitive processes in DSC remains a constructive challenge that probes – and ideally enhances – our understanding of human behavior in complex dynamic environments.

## AUTHOR CONTRIBUTIONS

Both authors contributed to writing the manuscript and approved it for publication.

## ACKNOWLEDGMENTS

We acknowledge the financial support of Deutsche Forschungsgemeinschaft (DFG) and Heidelberg University granted within the Open-Access Publishing Program.

## REFERENCES

- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psychol. Rev.* 111:1036. doi: 10.1037/0033-295X.111.4.1036
- Anzai, Y. (1984). Cognitive control of real-time event-driven systems. *Cogn. Sci.* 8, 221–254. doi: 10.1016/S0364-0213(84)80002-6
- Berry, D. C., and Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *Q. J. Exp. Psychol.* 36, 209–231. doi: 10.1080/14640748408402156
- Brehmer, B. (1992). Dynamic decision making: human control of complex systems. *Acta Psychol.* 81, 211–241. doi: 10.1016/0001-6918(92)90019-A
- Brehmer, B. (2005). Micro-worlds and the circular relation between people and their environment. *Theor. Issues Ergon. Sci.* 6, 73–93. doi: 10.1080/14639220512331311580
- Brehmer, B., and Dörner, D. (1993). Experiments with computer-simulated microworlds: escaping both the narrow straits of the laboratory and the deep blue sea of the field study. *Comput. Hum. Behav.* 9, 171–184. doi: 10.1016/0747-5632(93)90005-D
- Brehmer, B., and Elg, F. (2005). “Heuristics in dynamic decision making: coping with the time constants of a dynamic task by doing something else,” in *Proceedings of the 23rd International Conference of the System Dynamics Society*, eds J. D. Serman, N. P. Repenning, R. S. Langer, J. I. Rowe, and J. M. Yanni, Boston, MA.
- Broadbent, D. E., FitzGerald, P., and Broadbent, M. H. (1986). Implicit and explicit knowledge in the control of complex systems. *Br. J. Psychol.* 77, 33–50. doi: 10.1111/j.2044-8295.1986.tb01979.x
- Cronin, M. A., Gonzalez, C., and Serman, J. D. (2009). Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens. *Organ. Behav. Hum. Decis. Process.* 108, 116–130. doi: 10.1016/j.obhdp.2008.03.003
- Danner, D., Hagemann, D., Holt, D. V., Hager, M., Schankin, A., Wüstenberg, S., et al. (2011). Measuring performance in dynamic decision making: reliability and validity of the tailorshop simulation. *J. Individ. Differ.* 32, 225–233. doi: 10.1027/1614-0001/a000055
- De Obeso Orendain, A. D. O., and Wood, S. (2012). “An account of cognitive flexibility and inflexibility for a complex dynamic task,” in *Proceedings of the 11th International Conference on Cognitive Modeling*, eds N. Russwinkel, U. Drewitz, and H. van Rijn (Berlin: TU Berlin), 49–54.
- Dienes, Z., and Fahey, R. (1995). Role of specific instances in controlling a dynamic system. *J. Exp. Psychol.* 21:848. doi: 10.1037/0278-7393.21.4.848
- Dutt, V., and Gonzalez, C. (2012). Making instance-based learning theory usable and understandable: the instance-based learning tool. *Comput. Hum. Behav.* 28, 1227–1240. doi: 10.1016/j.chb.2012.02.006

- Edwards, W. (1962). Dynamic decision theory and probabilistic information processing. *Hum. Factors* 4, 59–74. doi: 10.1177/001872086200400201
- Erev, I., and Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychol. Rev.* 112:912. doi: 10.1037/0033-295X.112.4.912
- Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Think. Reason.* 7, 69–89. doi: 10.1080/13546780042000046
- Fum, D., and Stocco, A. (2003). “Instance vs. rule-based learning in controlling a dynamic system,” in *Proceedings of the 5th International Conference on Cognitive Modelling*, eds F. Detje, D. Dörner, and H. Schaub (Bamberg: Universitäts-Verlag), 105–110.
- Gibson, F. P., Fichman, M., and Plaut, D. C. (1997). Learning in dynamic decision tasks: computational model and empirical evidence. *Organ. Behav. Hum. Decis. Process.* 71, 1–35. doi: 10.1006/obhd.1997.2712
- Gigerenzer, G., and Brighton, H. (2009). Homo heuristics: why biased minds make better inferences. *Top. Cogn. Sci.* 1, 107–143. doi: 10.1111/j.1756-8765.2008.01006.x
- Gilovich, T., Griffin, D., and Kahneman, D. (eds) (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511808098
- Glass, B. D., and Osman, M. (2017). Positive explorers: modeling dynamic control in normal aging. *Aging Neuropsychol. Cogn.* 24, 62–79. doi: 10.1080/13825585.2016.1171290
- Gonzalez, C., Dutt, V., Healy, A. F., Young, M. D., and Bourne, L. E. Jr. (2009). “Comparison of instance and strategy models in ACT-R,” in *Proceedings of the 9th International Conference on Cognitive Modeling*, eds H. D. Peebles and R. Cooper (Manchester: University of Manchester).
- Gonzalez, C., Lerch, J. F., and Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cogn. Sci.* 27, 591–635. doi: 10.1016/S0364-0213(03)00031-4
- Gonzalez, C., Vanyukov, P., and Martin, M. K. (2005). The use of microworlds to study dynamic decision making. *Comput. Hum. Behav.* 21, 273–286. doi: 10.1016/j.chb.2004.02.014
- Grosskopf, B. (2003). Reinforcement and directional learning in the ultimatum game with responder competition. *Exp. Econ.* 6, 141–158. doi: 10.1023/A:10253050
- Gureckis, T. M., and Love, B. C. (2009). Learning in noise: dynamic decision-making in a variable environment. *J. Math. Psychol.* 53, 180–193. doi: 10.1016/j.jmp.2009.02.004
- Herde, C. N., Wüstenberg, S., and Greiff, S. (2016). Assessment of complex problem solving: what we know and what we don't know. *Appl. Meas. Educ.* 29, 265–277. doi: 10.1080/08957347.2016.1209208
- Holyoak, K. J., and Cheng, P. W. (2011). Causal learning and inference as a rational process: the new synthesis. *Annu. Rev. Psychol.* 62, 135–163. doi: 10.1146/annurev.psych.121208.131634
- Klein, G. (2002). “The fiction of optimization,” in *Bounded Rationality: The Adaptive Toolbox*, eds G. Gigerenzer and R. Selten (Cambridge, MA: MIT Press).
- Klein, G. A., Orasanu, J., Calderwood, R., and Zsombok, C. E. (1993). *Decision Making in Action: Models and Methods*. Norwood, NJ: Ablex.
- Kluge, A. (2008). Performance assessments with microworlds and their difficulty. *Appl. Psychol. Meas.* 32, 156–180. doi: 10.1177/0146621607300015
- Lewandowsky, S., and Farrell, S. (2011). *Computational Modeling in Cognition: Principles and Practice*. Thousand Oaks, CA: Sage. doi: 10.4135/9781483349428
- Marewski, J. N., and Mehlhorn, K. (2011). Using the ACT-R architecture to specify 39 quantitative process models of decision making. *Judgm. Decis. Mak.* 6, 439–519.
- Marewski, J. N., and Schooler, L. J. (2011). Cognitive niches: an ecological model of strategy selection. *Psychol. Rev.* 118, 393–437. doi: 10.1037/a0024143
- Meder, B., Gerstenberg, T., Hagmayer, Y., and Waldmann, M. R. (2010). Observing and intervening: rational and heuristic models of causal decision making. *Open Psychol. J.* 3, 119–135. doi: 10.2174/1874350101003010119
- Newell, A., and Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.
- OECD (2014). *PISA 2012 Results Creative Problem Solving: Students' Skills in Tackling Real-life Problems*, Vol. V. Paris: OECD. doi: 10.1787/9789264208070-en
- Osman, M. (2010). Controlling uncertainty: a review of human behavior in complex dynamic environments. *Psychol. Bull.* 136, 65–86. doi: 10.1037/a0017815
- Osman, M. (2017). “Planning and Control,” in *Oxford Handbook of Causal Reasoning*, ed. M. Waldmann (Oxford: Oxford University Press).
- Osman, M., Glass, B. D., and Holo, Z. (2015). Approaches to learning to control dynamic uncertainty. *Systems* 3, 211–236. doi: 10.3390/systems3040211
- Reichert, U., and Dörner, D. (1988). Heuristics beim Umgang mit einem “einfachen” dynamischem system [heuristics in handling a “simple” dynamic system]. *Spr. Kogn.* 7, 12–24.
- Sager, S., Barth, C. M., Diedam, H., Engelhart, M., and Funke, J. (2011). Optimization as an analysis tool for human complex problem solving. *SIAM J. Optim.* 21, 936–959. doi: 10.1137/11082018X
- Schoelles, M. J., and Gray, W. D. (2000). “Argus prime: modeling emergent microstrategies in a complex simulated task environment,” in *Proceedings of the 3rd International Conference on Cognitive Modeling*, eds N. Taatgen and J. Aasman (Veenendaal: Universal Press), 260–270.
- Schoppek, W. (2002). Examples, rules, and strategies in the control of dynamic systems. *Cogn. Sci. Q.* 2, 63–92.
- Schoppek, W., and Boehm-Davis, D. A. (2004). Opportunities and challenges of modeling user behavior in complex real world tasks. *MMI Interaktiv* 7, 47–60.
- Schoppek, W., and Fischer, A. (2015). Complex problem solving—single ability or complex phenomenon? *Front. Psychol.* 6:1669. doi: 10.3389/fpsyg.2015.01669
- Schunn, C. D., and Anderson, J. R. (1998). “Scientific discovery,” in *The Atomic Components of Thought*, eds J. R. Anderson and C. Lebiere (New York, NY: Psychology Press), 385–427.
- Selten, R., and Stoeker, R. (1986). End behavior in sequences of finite Prisoner's Dilemma supergames A learning theory approach. *J. Econ. Behav. Organ.* 7, 47–70. doi: 10.1016/0167-2681(86)90021-1
- Shah, A. K., and Oppenheimer, D. M. (2008). Heuristics made easy: an effort-reduction framework. *Psychol. Bull.* 134:207. doi: 10.1037/0033-2909.134.2.207
- Sterman, J. D. (1989). Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment. *Manag. Sci.* 35, 321–339. doi: 10.1287/mnsc.35.3.321
- Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., and Blum, B. (2003). Inferring causal networks from observations and interventions. *Cogn. Sci.* 27, 453–489. doi: 10.1016/S0364-0213(03)00010-7
- Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.
- Taatgen, N., and Anderson, J. R. (2010). The past, present, and future of cognitive architectures. *Top. Cogn. Sci.* 2, 693–704. doi: 10.1111/j.1756-8765.2009.01063.x
- Taatgen, N. A., and Wallach, D. (2002). Whether skill acquisition is rule or instance based is determined by the structure of the task. *Cogn. Sci. Q.* 2, 163–204.
- Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131. doi: 10.1126/science.185.4157.1124
- Wüstenberg, S., Greiff, S., and Funke, J. (2012). Complex problem solving—More than reasoning? *Intelligence* 40, 1–14. doi: 10.1016/j.intell.2011.11.003

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Holt and Osman. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Evaluation-Dependent Representation in Risk Defusing

Oswald Huber\*

Department of Psychology, University of Fribourg, Fribourg, Switzerland

**Keywords:** decision making, risky decision making, representation, evaluation, risk defusing, risk defusing operator (RDO), advantages first principle

This paper treats the effect of the evaluation of outcomes on the representation in a decision process. I assume that how the outcomes are evaluated up to a specific step affects the representation in this step. Thus, representation and evaluation in the process are intermingled.

Research in the tradition of Psychological decision theory investigates risky decisions in experiments generally using gambles as alternatives, or alternatives that are designed like gambles by the experimenter. A gamble is characterized by its outcomes (gains, losses) and their probabilities, all these are known to the decision maker. The fundamental influences determining decision behavior in such experiments are the subjective values (utilities) of the outcomes and frequently their subjective probability. The most prominent decision theories founded in the gambling paradigm are Subjectively Expected Utility theory and its descendants, e.g., Prospect theory Kahneman and Tversky (1979), Baron (2008) gives an overview. Also, decision heuristics are based on one or more of these components (Shah and Oppenheimer, 2008). Most papers in this review paper are based on process tracing methods (e.g., alternatives  $\times$  dimensions-matrix, verbal protocols).

If in experiments realistic scenarios are used instead of gambles, decision behavior differs in two respects: First, decision makers are often content with knowing whether a certain outcome occurs with certainty or is possible, and are not actively interested in more precise probabilities. Second, decision makers often actively seek a risk defusing operator which reduces the risk.

A *risk-defusing operator* (RDO) is an action that is anticipated to remove or reduce the risk. This action is planned by the decision maker to be carried out *in addition* to an existing alternative. Consider, for example, the situation of a person who thinks about the action alternative to travel into a country where a contagious illness endures (Huber, 2012). This person may inquire, for example, whether a vaccination exists. Getting vaccinated is an RDO preventing the negative outcome (infection). If a satisfactory RDO is found with an otherwise attractive alternative, this alternative is usually chosen (e.g., Bär and Huber, 2008). An overview about the decision process and experimental results concerning RDOs is presented in Huber (2007, 2012).

There are various types of RDOs (e.g., Huber, 2007): one type *prevents* a negative outcome (e.g., vaccination, drinking bottled water only), for example, or another *compensates* for a negative outcome (e.g., insurance). Search for an RDO often involves a price (money, time, effort, etc.), and at the start it is unclear whether the search turns out to be effective. Understandably, search is more likely if the expectation of success is higher (Huber and Huber, 2008). Search is also more likely under time pressure (Huber and Kunz, 2007) and under justification pressure (Huber et al., 2009). Moreover, the type of risk influences the search (Wilde et al., 2008). An RDO is not automatically satisfactory: The higher the cost, the less likely is the RDO accepted (Williamson et al., 2000; Huber and Huber, 2003). In a multistage investment task too, people are willing to purchase an RDO, if they are given the opportunity (Huber, 1996).

The conception of RDOs is overlooked in classical decision research. This disregard seems to result from the belief that *all* risky decisions involve gambles, where RDOs are not relevant. Gambles can be considered to form only a subclass of risky decision tasks.

## OPEN ACCESS

### Edited by:

Joachim Funke,  
Heidelberg University, Germany

### Reviewed by:

Anton Kühberger,  
University of Salzburg, Austria  
Andreas Fischer,  
Forschungsinstitut Betriebliche  
Bildung, Germany

### \*Correspondence:

Oswald Huber  
oswald.huber@unifr.ch

### Specialty section:

This article was submitted to  
Cognitive Science,  
a section of the journal  
Frontiers in Psychology

**Received:** 21 February 2017

**Accepted:** 08 May 2017

**Published:** 23 May 2017

### Citation:

Huber O (2017) Evaluation-Dependent  
Representation in Risk Defusing.  
Front. Psychol. 8:836.  
doi: 10.3389/fpsyg.2017.00836

## DYNAMIC MENTAL REPRESENTATION OF ALTERNATIVES

Classical descriptive decision theory is based on Subjective Expected Utility Theory and its modifications. In these theories, the representation of alternatives is rarely addressed explicitly. Prospect theory (Kahneman and Tversky, 1979) is an exception. The authors suppose that representation is a distinct first phase of the decision process, and then—in a second phase—the alternatives are evaluated.

In the risk-defusing context, a different approach is taken. Decision makers in non-routine situations are assumed to construct a mental representation by sequentially incorporating new information they consider as relevant (outcomes, probabilities, RDOs, ...) into a causal mental model of the alternatives (Huber, 2011). The representation is *dynamic*, not only because it is constructed in time, but also because it may be changed (by introducing an RDO) in the course of elaboration. These mental models usually do include a different amount of elements for different alternatives: more information is represented for some alternatives, less for others. I assume that normally items like outcomes are evaluated immediately when they are introduced into the representation as more or less desirable or undesirable. How much is represented for each alternative, depends on this evaluation, as described in the next section. Thus, elaboration and evaluative processes are closely intermingled (in contrast to, e.g., Prospect theory).

## ADVANTAGES FIRST PRINCIPLE

We have shown in three experiments that the great majority of decision makers follow the Advantages first Principle (Huber et al., 2011). This principle describes, how information search is guided by the evaluation up to now:

1. Decision makers consider mainly outcomes evaluated as positive, and based on this evaluation, promising alternatives are distinguished from not promising ones.
2. They then inspect the promising alternatives further and deliberately search—among others—for undesirable outcomes<sup>1</sup>. Of course, when they detect a negative outcome they may search for an RDO.

Thus, the Advantages first Principle is not a choice heuristic (selecting *the* subjectively best alternative) but a heuristic to select promising problem solving paths. The situation is similar to chess, where expert players do not examine every possible move. Instead, they center on few moves that seem worth pursuing founded on a preliminary evaluation (Holding, 1992). The early search for positive or negative outcomes when decision makers have no previous information has not been studied systematically in decision research, to the best of my knowledge.

The decision maker concentrates on positive outcomes and attractive alternatives because elaboration is expensive in time, effort, etc., and he or she wants to reduce these cost. (1) If an alternative has an attractive positive outcome, time and effort

can be invested into this alternative to examine it in more detail. A negative outcome may be defused with an acceptable RDO. (2) If, however, the positive outcome of an alternative is only mediocre, then it remains only mediocre, even though no negative outcome should turn up. (3) Concentrating initially on the negative outcomes would not be an economical heuristic to decide which alternatives to inspect more. An exception is an alternative with a very negative outcome which cannot be defused. In this case, it can be ignored and no information has to be searched for. In all other situations, it would always be essential to also check the positive outcomes. Otherwise, one might invest much time and effort in an alternative that later turns out to be inferior.

To recapitulate, the Advantages first Principle defines a rational heuristic for selecting alternatives deserving a more careful inspection. The decision maker can generally at all times return to an alternative that he or she thinks should have been examined deeper.

As mentioned above, the majority of decision makers uses the Advances first Principle (more than 80%), but a minority investigated negative outcomes first. Huber et al. (2011) presume that the principle is applied, when (a) decision makers are free to acquire the information in the order they prefer, (b) they do not already have exhaustive knowledge about the alternatives (they are no experts), (c) they do not assume that the set of available alternatives is a (positive) pre-selection with good positive outcomes for all alternatives, and (d) there is no acceptance criterion that all alternatives have to fulfill (like the maximum rent). Furthermore, in time pressure conditions, most people start with a negative outcome, but without time pressure, they start with a positive one (Huber and Kunz, 2007). Time pressure here means that there is a kind of (external or internal) deadline *and* the decision maker realizes that the available time may be too short to make a decision (e.g., Benson and Beach, 1996).

The Advantages first Principle is, as stated above, a heuristic for selecting a promising problem solving path. So possibly an alternative picked out first is not chosen, e.g., because a negative outcome cannot be defused. This is not a falsifying instance for Advantages first, because it does not predict choices. There should, though, be a correlation between selecting a promising alternative and the chosen one, albeit we do not know at present how big this correlation is. Therefore, it will be essential to investigate the research that found negative outcomes to have a more pronounced effect, as, for example, the framing effect for gains (Tversky and Kahneman, 1981), or Priority heuristic (Brandstätter et al., 2006).

## REPRESENTATION—POSITIVE AND NEGATIVE OUTCOMES

Up to now, we have only considered RDOs defusing a negative outcome (*negative* RDOs). An RDO may, however, improve the chance to receive a positive outcome (*positive* RDOs). An example is doping in sports, which increases the chances of winning. The question I want to address here is whether positive RDOs are searched for equally often as negative ones.

<sup>1</sup>This result speaks against Montgomery and Willén (1999) *dominance structuring* model. For a more detailed discussion, see Huber (2011).



We investigated this question by embedding the decisions in a framing context, see Huber et al. (2014) for details. Otherwise we are confronted with the problem that frames causing search for either positive or for negative RDOs often involve outcomes with distinct attractiveness, and—as the previous section clarified—attractiveness is one of the central factors influencing RDO search. The result was clear: Much more people (83%) searched for a negative RDO, whereas only 39% searched for a positive one. We attribute this result to a general readiness for negative stimuli and for starting adequate reactions (Taylor, 1991; Öhman et al., 2000; Woody and Szechtman, 2013). Such readiness is favorable because if somebody is provoked with a possible danger (e.g., a person aiming a pistol on me, or a snake), it is frequently vital to react fast. Positive stimuli are not connected with an analogous general readiness, at least nothing comparable is reported in the literature. Searching for an RDO preventing or compensating the negative stimulus clearly is such an adequate reaction when confronted with a possible negative outcome.

Thus, negative RDOs are activated or investigated more often when otherwise a negative alteration is possible than positive ones when a positive alteration is possible. So, in this case too, evaluation of the alternatives is a factor determining the construction of a representation.

## CONCLUSION

The previous sections have demonstrated that how the outcomes are evaluated is a central aspect in the process of constructing

a representation. I want to empathize I expect the Advantages First Principle to hold in what we could call non-routine situations, as described in the relevant section. The Advantages first Principle is explicitly not a choice heuristic, and does not predict that choice is based solely on positive outcomes. It is, to repeat, a heuristic selecting a path that is worthwhile to be followed.

I could not go into details of theories that would be useful to be investigated, for example, Naturalistic decision making and Query theory. Naturalistic decision making (Klein, 1999) deals with realistic decision situations. It is, however, concerned mainly with non-experimental research in decisions of experts, and experts are explicitly not the topic of my paper. Query theory (e.g., Johnson et al., 2007) proposes people to construct their values. The used method is interesting and seems to be a variant of thinking aloud.

By concentrating on evaluation I did of course not want to exclude other influences on risk defusing. Such effects are, for example: justification pressure, the expectation of search success, the type of risk involved, or the expectation to get useful probability information. However, an inclusion of these topics is beyond the scope of this paper.

## AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

## REFERENCES

- Bär, A. S., and Huber, O. (2008). Successful or unsuccessful search for risk defusing operators: effect on decision behavior. *Eur. J. Cogn. Psychol.* 20, 807–827. doi: 10.1080/09541440701686227
- Baron, J. (2008). *Thinking and Deciding, 4th Edn.* Cambridge, NY: Cambridge University Press.
- Benson, L., and Beach, L. R. (1996). The effects of time constraints on the pre-choice screening of decision options. *Organ. Behav. Hum. Decis. Process.* 67, 222–228. doi: 10.1006/obhd.1996.0075
- Brandstätter, E., Gigerenzer, G., and Hertwig, R. (2006). The priority heuristic: making choices without trade-offs. *Psychol. Rev.* 113, 409–431. doi: 10.1037/0033-295X.113.2.409
- Holding, D. H. (1992). Theories of chess skill. *Psychol. Res.* 54, 10–16. doi: 10.1007/BF01359218
- Huber, O. (1996). Buying protection in a multistage investment task. *Acta Psychol.* 92, 153–167. doi: 10.1016/0001-6918(95)00009-7
- Huber, O. (2007). “Behavior in risky decisions: focus on risk defusing,” in *Uncertainty and Risk*, eds M. Abdellaoui, R. D. Luce, M. Machina, and B. Munier (Berlin; New York, NY: Springer), 291–306.
- Huber, O. (2011). “Risky decision making with Mental Causal Models,” in *Perspectives on Thinking, Judging, and Decision Making*, eds W. Brun, G. Keren, G. Kirkeboen, and H. Montgomery (Oslo: Universitetsforlaget), 179–189.
- Huber, O. (2012). Risky decisions: active risk management. *Curr. Dir. Psychol. Sci.* 21, 26–30. doi: 10.1177/096372141142205
- Huber, O., Bär, A. S., and Huber, O. W. (2009). Justification pressure in risky decision making: search for risk defusing operators. *Acta Psychol.* 130, 17–24. doi: 10.1016/j.actpsy.2008.09.009
- Huber, O., and Huber, O. W. (2003). Detectability of the negative event: effect on the acceptance of pre- or post-event risk-defusing actions. *Acta Psychol.* 113, 1–21. doi: 10.1016/S0001-6918(02)00148-8
- Huber, O., and Huber, O. W. (2008). Gambles versus quasi-realistic scenarios: expectations to find probability- and risk defusing information. *Acta Psychol.* 127, 222–236. doi: 10.1016/j.actpsy.2007.05.002
- Huber, O., Huber, O. W., and Bär, A. S. (2011). Information search and mental representation in risky decision making: the advantages first principle. *J. Behav. Decis. Mak.* 24, 223–248. doi: 10.1002/bdm.674
- Huber, O., Huber, O. W., and Bär, A. S. (2014). Framing of decisions: effect on active and passive risk avoidance. *J. Behav. Decis. Mak.* 27, 444–453. doi: 10.1002/bdm.1821
- Huber, O., and Kunz, U. (2007). Time pressure in risky decision making: effect on risk defusing. *Psychol. Sci.* 49, 415–426.
- Johnson, E. J., Häubl, G., and Keinan, A. (2007). Aspects of endowment: a query theory of value construction. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 461–474. doi: 10.1037/0278-7393.33.3.461
- Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–291. doi: 10.2307/1914185
- Klein, G. A. (1999). *Sources of Power: How People Make Decisions.* Cambridge, MA: MIT Press.
- Montgomery, H., and Willén, H. (1999). “Decision making and action: the search for a good structure,” in *Judgment and Decision Making*, eds P. Juslin and H. Montgomery (Mahwah, NJ: Lawrence Erlbaum), 147–173.
- Öhman, A., Flykt, A., and Lundqvist, D. (2000). “Unconscious emotion: evolutionary perspectives, psychophysiological data and neuropsychological mechanisms,” in *Cognitive Neuroscience of Emotion*, eds R. D. Lane, L. Nadel and G. L. Ahern (New York, NY: Oxford University Press), 296–327.
- Shah, A. K., and Oppenheimer, D. M. (2008). Heuristics made easy: an effort-reduction framework. *Psychol. Bull.* 134, 207–222. doi: 10.1037/0033-2909.134.2.207
- Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: the mobilization-minimization hypothesis. *Psychol. Bull.* 110, 67–85. doi: 10.1037/0033-2909.110.1.67

- Tversky, A., and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science* 211, 453–458. doi: 10.1126/science.7455683
- Wilke, M., Haug, H., and Funke, J. (2008). Risk-specific search for risk-defusing operators. *Swiss J. Psychol.* 67, 29–40. doi: 10.1024/1421-0185.67.1.29
- Williamson, J., Ranyard, R., and Cuthbert, L. (2000). Risk management in naturalistic insurance decisions: evidence from a simulation study. *Risk Decis. Policy* 5, 19–38. doi: 10.1017/S1357530900000090
- Woody, E. Z., and Szechtman, H. (2013). A biological security motivation system for potential threats: are there implications for policy-making? *Front. Hum. Neurosci.* 7:556. doi: 10.3389/fnhum.2013.00556

**Conflict of Interest Statement:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Huber. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Impact of Cognitive Abilities and Prior Knowledge on Complex Problem Solving Performance – Empirical Results and a Plea for Ecologically Valid Microworlds

Heinz-Martin Süß<sup>1\*</sup> and André Kretschmar<sup>2</sup>

<sup>1</sup> Institute of Psychology, Otto-von-Guericke University Magdeburg, Magdeburg, Germany, <sup>2</sup> Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany

## OPEN ACCESS

### Edited by:

Wolfgang Schoppek,  
University of Bayreuth, Germany

### Reviewed by:

Natassia Goode,  
University of the Sunshine Coast,  
Australia  
Marc Halbrügge,  
Technische Universität Berlin,  
Germany

### \*Correspondence:

Heinz-Martin Süß  
heinz-martin.suess@ovgu.de

### Specialty section:

This article was submitted to  
Cognitive Science,  
a section of the journal  
Frontiers in Psychology

**Received:** 06 October 2017

**Accepted:** 13 April 2018

**Published:** 08 May 2018

### Citation:

Süß H-M and Kretschmar A (2018)  
Impact of Cognitive Abilities and Prior  
Knowledge on Complex Problem  
Solving Performance – Empirical  
Results and a Plea for Ecologically  
Valid Microworlds.  
Front. Psychol. 9:626.  
doi: 10.3389/fpsyg.2018.00626

The original aim of complex problem solving (CPS) research was to bring the cognitive demands of complex real-life problems into the lab in order to investigate problem solving behavior and performance under controlled conditions. Up until now, the validity of psychometric intelligence constructs has been scrutinized with regard to its importance for CPS performance. At the same time, different CPS measurement approaches competing for the title of the best way to assess CPS have been developed. In the first part of the paper, we investigate the predictability of CPS performance on the basis of the Berlin Intelligence Structure Model and Cattell's investment theory as well as an elaborated knowledge taxonomy. In the first study, 137 students managed a simulated shirt factory (*Tailorshop*; i.e., a complex real life-oriented system) twice, while in the second study, 152 students completed a forestry scenario (*FSYS*; i.e., a complex artificial world system). The results indicate that reasoning – specifically numerical reasoning (Studies 1 and 2) and figural reasoning (Study 2) – are the only relevant predictors among the intelligence constructs. We discuss the results with reference to the Brunswik symmetry principle. Path models suggest that reasoning and prior knowledge influence problem solving performance in the *Tailorshop* scenario mainly indirectly. In addition, different types of system-specific knowledge independently contribute to predicting CPS performance. The results of Study 2 indicate that working memory capacity, assessed as an additional predictor, has no incremental validity beyond reasoning. We conclude that (1) cognitive abilities and prior knowledge are substantial predictors of CPS performance, and (2) in contrast to former and recent interpretations, there is insufficient evidence to consider CPS a unique ability construct. In the second part of the paper, we discuss our results in light of recent CPS research, which predominantly utilizes the minimally complex systems (MCS) measurement approach. We suggest ecologically valid microworlds as an indispensable tool for future CPS research and applications.

**Keywords:** complex problem solving, microworlds, minimally complex systems, intelligence, investment theory, knowledge assessment, working memory, Brunswik symmetry

## INTRODUCTION

People are frequently confronted with problems in their daily lives that can be characterized as complex in many aspects. A subset of these problems can be described as interactions between a person and a dynamic system of interconnected variables. By manipulating some of these variables, the person can try to move the system from its present state to a goal state or keep certain critical variables within tolerable ranges. Problems of this kind can be simulated using computer models (aka microworlds), offering an opportunity to observe human behavior in realistic problem environments under controlled conditions.

The study of human interaction with complex computer-simulated problem scenarios has become an increasingly popular field of research in numerous areas of psychology over the past four decades. For example, computer models have been built to simulate the job of a small-town mayor (Dörner et al., 1983), a production plant operator (Bainbridge, 1974; Morris and Rouse, 1985), a business manager (Putz-Osterloh, 1981; Wolfe and Roberts, 1986), a coal-fired power plant operator (Wallach, 1997), and a water distribution system operator (Gonzalez et al., 2003). Real-time simulations have put users in the role of the head of a firefighting crew (Brehmer, 1986; Rigas et al., 2002) or an air traffic controller (Ackerman and Kanfer, 1993). In experimental psychology, research on complex problem solving (CPS) has sought to formally describe simulations (e.g., Buchner and Funke, 1993; Funke, 1993), the effects of system features on task difficulty (e.g., Funke, 1985; Gonzalez and Dutt, 2011), the role of emotions (e.g., Spering et al., 2005; Barth and Funke, 2010), and the effects of practice and training programs (e.g., Kluge, 2008b; Kretzschmar and Süß, 2015; Goode and Beckmann, 2016; Engelhart et al., 2017; see also Funke, 1995, 1998). Differential and cognitive psychology research has investigated the psychometrical features of CPS assessments (e.g., Rigas et al., 2002), the utility of computational models for explaining CPS performance (e.g., Dutt and Gonzalez, 2015), the relationship between CPS performance and cognitive abilities (e.g., Wittmann and Süß, 1999), and its ability to predict real-life success criteria (e.g., Kersting, 2001). For detailed summaries of different areas of CPS research, see Frensch and Funke (1995) and Funke (2006).

Meanwhile, many researchers have moved away from complex real life-oriented systems (CRS) to complex artificial world systems (CAS) in order to increase the psychometric quality of measures and to control for the effects of preexisting knowledge (e.g., Funke, 1992; Wagener, 2001; Kröner et al., 2005). This development ultimately culminated in the minimally complex systems (MCS) approach (Greiff et al., 2012), also known as the multiple complex systems approach (e.g., Greiff et al., 2015a). This approach has recently become prominent in educational psychology (e.g., Greiff et al., 2013b; Sonleitner et al., 2013; Kretzschmar et al., 2014; OECD, 2014; Csapó and Molnár, 2017). In addition, this shift has led to the question of what are and are not complex problems, with some researchers questioning the relevance of MCS as a tool for CPS research and the validity of the conclusions drawn from them (e.g., Funke, 2014; Dörner and Funke, 2017; Funke et al., 2017; Kretzschmar, 2017).

Originally, simulated dynamic task environments were used to reproduce the cognitive demands associated with real-life problems in the laboratory (Dörner et al., 1983; Dörner, 1986). These environments have several features: (1) Complexity: Many aspects of a situation must be taken into account at the same time. (2) Interconnectivity: The different aspects of a situation are not independent of one another and therefore cannot be controlled separately. (3) Intransparency: Only some of the relevant information is made available to the problem solver. (4) Dynamics: Changes in the system occur without intervention from the agent. (5) Polytele: The problem solver must sometimes pursue multiple and even contradictory goals simultaneously. (6) Vagueness: Goals are only vaguely formulated and must be defined more precisely by the problem solver. Whereas older microworlds featured all of these characteristics to a considerable extent, more recent approaches such as MCS have substituted complexity and ecological validity (i.e., the simulation's validity as a realistic problem-solving environment allowing psychological statements to be made about the real world; see Fahrenberg, 2017) for highly reliable assessment instruments by simulating tiny artificial world relationships (e.g., Greiff et al., 2012; Sonleitner et al., 2012).

The present paper is divided into two parts. In the first part, we deal with one of the oldest but still an ongoing issue in the area of CPS research: the cognitive prerequisites of CPS performance. In two different studies, we used microworlds (CRS and CAS) to empirically investigate the impact of cognitive abilities (i.e., intelligence and working memory capacity) and prior knowledge on CPS performance. In doing so, we considered the impact of the Brunswik symmetry principle, which effects the empirical correlations between hierarchical constructs (e.g., Wittmann, 1988). Integrating our results with previous CPS research, we review the basis and empirical evidence for 'complex problem solving ability' as a distinct cognitive construct. In the second part of the paper, we discuss our approach and results in light of recent problem solving research, which predominantly utilizes the MCS approach. Finally, we conclude with some recommendations for future research on CPS and suggest ecologically valid microworlds as tools for research and applications.

## PART I: EMPIRICAL INVESTIGATION OF THE COGNITIVE PREREQUISITES OF COMPLEX PROBLEM SOLVING PERFORMANCE

### Intelligence and Complex Problem Solving

At the beginning of complex problem solving (CPS) research, CPS pioneers raised sharp criticisms of the validity of psychometric intelligence tests (Putz-Osterloh, 1981; Dörner et al., 1983; Dörner and Kreuzig, 1983). These measures, derisively referred to as "test intelligence," are argued to be bad predictors of performance on partially intransparent, ill-defined complex problems. In contrast to simulated scenarios, intelligence test tasks are less complex, static, transparent,



and well-defined problems that do not resemble most real-life demands in any relevant way. Zero correlations between intelligence measures and CPS performance were interpreted as evidence of the discriminant validity of CPS assessments, leading to the development of a new ability construct labeled complex problem solving ability or operative intelligence (Dörner, 1986). However, no evidence of the convergent validity of CPS assessments or empirical evidence for their predictive validity with regard to relevant external criteria or even incremental validity beyond psychometric intelligence tests have been presented.

By now, numerous studies have investigated the relationship between control performance on computer-simulated complex systems and intelligence. Whereas Kluwe et al. (1991) found no evidence of a relationship in an older review, more recent studies have found correlations that are substantial but still modest enough to argue in favor of a distinct CPS construct (e.g., Wüstenberg et al., 2012; Greiff et al., 2013b; Sonleitner et al., 2013). In a more recent meta-analysis, Stadler et al. (2015) calculated the overall average effect size between general intelligence ( $g$ ) and CPS performance to be  $r = 0.43$  (excluding outliers,  $r = 0.40$ ), with a 95% confidence interval ranging from 0.37 to 0.49. The mean correlation between CPS performance and reasoning was  $r = 0.47$  (95% CI: 0.40 to 0.54). The relationship with  $g$  was stronger for MCS ( $r = 0.58$ ) than CRSs ( $r = 0.34$ )<sup>1</sup>. From our point of view, this difference results from the higher reliability of MCS but also a difference in cognitive demands. MCS are tiny artificial world simulations in which domain-specific prior knowledge is irrelevant. Complex real life-oriented tasks, however, activate preexisting knowledge about the simulated domain. This knowledge facilitates problem solving; in some cases, the problems are so complex that they cannot be solved at all without prior knowledge (e.g., Hesse, 1982).

The main issues with many complex real life-oriented studies that investigated the relation between intelligence and CPS performance concern the ecological validity of the simulations and the psychometric quality of the problem-solving performance criteria. This often leads to much larger confidence intervals in their correlations with intelligence compared to minimal complex tasks (Stadler et al., 2015). When the goals of a simulation are multiple and vaguely defined, the validity of any objective criterion is questionable since it might not correspond to the problem solver's subjective goal. However, people are unlikely to face a single, well-defined goal in real-life problems, limiting the ecological validity of such systems – despite the fact that a well-defined goal is a necessary precondition for assessing problem solving success in a standardized way, which is necessary in order to compare subjects' performance. Moreover, single problem solving trials produce only “single act criteria” (Fishbein and Ajzen, 1974), criticized as “one-item-testing” (e.g., Wüstenberg et al., 2012), the reliability of which is severely limited. Performance scores must be aggregated via repeated measurements to increase the proportion of reliable

variance that can be predicted (e.g., Wittmann and Süß, 1999; Rigas et al., 2002). The MCS has implemented these steps, resulting in strong reliability estimates (e.g., Greiff et al., 2012; Sonleitner et al., 2012).

Another crucial issue with regard to the relation between intelligence and CPS performance is the operationalization of intelligence. Numerous prior studies have used a measure of general intelligence ( $g$ ) to predict problem solving success. Since  $g$  is a compound of several more specific abilities,  $g$  scores comprise variance in abilities relevant to complex problem solving as well as variance in irrelevant abilities. According to Wittmann's (1988) multivariate reliability theory and the Brunswik symmetry principle (see also Wittmann and Süß, 1999), this results in an asymmetric relationship between predictor and criterion, attenuating their correlation. More specific subconstructs of intelligence might be more symmetrical predictors because they exclude irrelevant variance. In our view, controlling complex systems requires a great deal of reasoning ability (e.g., Süß, 1996; Wittmann and Süß, 1999; Kröner et al., 2005; Sonleitner et al., 2013; Kretzschmar et al., 2016, 2017). Inductive reasoning is required to detect systematic patterns within the ever-changing system states and develop viable hypotheses about the system's causal structure. Deductive reasoning is necessary to infer expectations about future developments from knowledge of causal connections and deduce more specific goals from higher-order goals. Abilities such as perceptual speed (except in real-time simulations), memory, and verbal fluency, meanwhile, should be less relevant for success in complex problem solving. In this sense, it is an open question in CPS research whether WMC, as a more basic ability construct (e.g., Süß et al., 2002; Oberauer et al., 2008), is a more symmetrical predictor of CPS performance than reasoning (for an overview of previous findings, see Zech et al., 2017).

In summary, a substantial correlation between intelligence and CPS performance measured with real life-oriented microworlds can be expected if (1) sufficient reliability of the CPS measures is ensured (e.g., aggregation via repeated measures), and (2) the best symmetrical intelligence construct is used (e.g., reasoning instead of general intelligence or perceptual speed).

## Knowledge and Complex Problem Solving

In addition to the debate about intelligence's contribution to complex problem solving, many researchers have pointed out the significance of knowledge for the successful control of complex systems (e.g., Bainbridge, 1974; Dörner et al., 1983; Chi et al., 1988; Goode and Beckmann, 2010; Beckmann and Goode, 2014). Expert knowledge is sometimes claimed to be the only important predictor of real-life problem solving success (Ceci and Liker, 1986), while others point out that both intelligence and knowledge contribute substantially to predicting job performance (Schmidt, 1992), which certainly includes complex problem solving.

Scenarios that accurately simulate real-world relationships provide an opportunity to draw on preexisting knowledge about the part of reality being simulated. That being said, a

<sup>1</sup> The correlation between complex real life-oriented systems and reasoning was not reported, nor was the effect of outliers on relationships other than that between CPS and  $g$ .

simulation never is exactly equivalent to what the problem solver has experienced before. Experts in a domain can make use of their knowledge to operate a simulation within that domain, but they are not automatically experts in the simulated scenario. The application of domain knowledge to the simulation requires a considerable amount of transfer. Following Cattell's investment theory (Cattell, 1987), we assume that intelligence, and particularly reasoning, plays an important role in mediating this transfer. Therefore both, intellectual abilities, particularly reasoning and prior knowledge of the simulated domain, should be powerful predictors of complex problem solving success, although the effect of intelligence has been found to be mainly indirect, mediated through knowledge (Schmidt et al., 1986; Schmidt, 1992).

The knowledge relevant for successfully controlling a complex system can be differentiated conceptually on two dimensions. First, knowledge about the system can be distinguished from knowledge about appropriate actions. System knowledge is knowledge about the features and structure of a system, such as what variables it consists of, how these variables are related, and what kind of behaviors the system tends to exhibit. Action-related knowledge is knowledge about what to do in order to pursue a given goal. In contrast to system knowledge, action knowledge is always bound to a specific goal. Studies by Vollmeyer et al. (1996) provided evidence for the distinction between system knowledge and action knowledge: Participants who acquired knowledge about a system during an exploration phase with or without a given goal performed equally well on a subsequent test trial with the same goal. However, the group which had not been given a specific goal during the exploration phase outperformed the group with the specific goal on a test with a new goal. Presumably, the specific goal group had learned mainly action knowledge, whereas the other group had acquired more system knowledge, which was then transferable to new goals.

A second distinction, independent of the first, exists between declarative and procedural knowledge. Declarative knowledge is knowledge that a person can represent symbolically in some way – verbally, graphically or otherwise. Declarative knowledge can be expressed as accurate answers to questions. Procedural knowledge, on the other hand, can be expressed only through accurate performance. The distinction between declarative and procedural knowledge is based on the conceptual difference between “knowing that” and “knowing how” (Ryle, 1949).

While system knowledge and action knowledge differ in content, declarative and procedural knowledge are different forms of knowledge. Therefore, the two dimensions can be conceived of as orthogonal. System knowledge and action knowledge can both be declarative: A person can talk about which variables are causally related to which other variables, but also about what to do in order to keep the system stable. Similarly, both system knowledge and action knowledge can also be procedural: Knowing how to stabilize a system without being able to express it is procedural action knowledge. Being able to mentally simulate a system or diagnose what variable is causing a disturbance without being able to give a full verbal account of the reasons is indicative of procedural system knowledge. Several

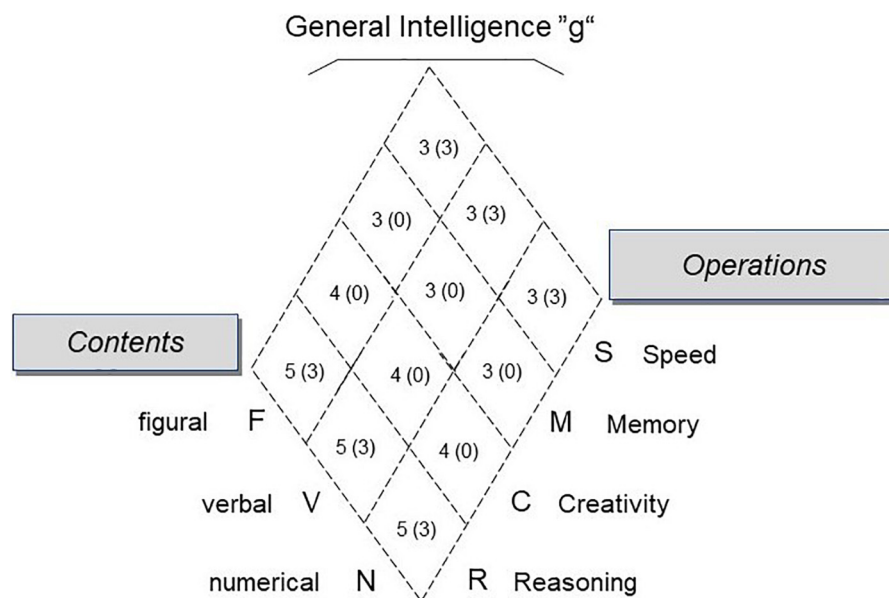
studies have found that people do not improve their problem-solving performance in controlling or repairing complex systems after receiving instructions in the form of declarative system knowledge (e.g., Morris and Rouse, 1985; Kluge, 2008b; but see Goode and Beckmann, 2010), and declarative knowledge sometimes is not correlated with problem solving performance (e.g., Berry and Dienes, 1993). Therefore, we must consider the possibility that procedural knowledge is part of the relevant knowledge base that guides a person's actions within complex dynamic environments.

In summary, prior domain knowledge must be considered as an additional substantial predictor of CPS performance. However, differentiating between different types of knowledge is necessary in order to explain CPS performance. In addition, different semantic embeddings (i.e., CRS vs. CAS) have different demands with regard to preexisting knowledge.

## The Present Study

The first goal of the two studies presented in this paper was to test the hypothesized criterion validity of reasoning in predicting problem solving performance in complex dynamic tasks. In addition, considering the Brunswik symmetry principle (Wittmann, 1988), we explored the predictive validity of additional more specific or more general intelligence constructs. Our investigation was based on the Berlin Intelligence Structure Model (BIS), a hierarchical and faceted model of intelligence (Jäger, 1982, 1984; for a detailed description in English, see Süß and Beauducel, 2015). The BIS differentiates intellectual abilities along two facets. The operation facet comprises four abilities: Reasoning (R) includes inductive, deductive and spatial reasoning and is equivalent to fluid intelligence (Gf). Creativity (C) refers to the ability to fluently produce many different ideas. Memory (M) refers to the ability to recall lists and configurations of items a few minutes after having learned them (episodic memory), whereas speed (S) refers to the ability to perform simple tasks quickly and accurately (perceptual speed). The second facet is postulated to include three content-related abilities: verbal (V), numerical (N) and figural-spatial (F) intelligence. Cross-classifying the four operational and three content abilities results in 12 lower-order cells. In addition, general intelligence is conceptualized as an overarching factor (Figure 1). For summaries of the validity and scope of the BIS, see the handbook for the BIS Test (Jäger et al., 1997) as well as Süß and Beauducel (2005, 2015).

In the second study, we included WMC as an additional predictor. Working memory is considered the most important cognitive resource for complex information processing, which includes reasoning (e.g., Kyllonen and Christal, 1990; Süß et al., 2002; Conway et al., 2003), language comprehension (e.g., King and Just, 1991), and math performance (e.g., Swanson and Kim, 2007). Consequently, previous research has found a significant relation between WMC and CPS (e.g., Wittmann and Süß, 1999; Bühner et al., 2008; Schweizer et al., 2013; Greiff et al., 2016). However, whether the more basic construct (i.e., WMC) is a stronger symmetrical predictor of CPS than reasoning from the perspective of the Brunswik symmetry principle (Wittmann, 1988) is not clear (for an overview, see Zech et al., 2017).



**FIGURE 1 |** The Berlin Intelligence Structure Model (BIS), and the number of tasks for each cell applied in Study 1 (in brackets, Study 2). In the BIS, four operation ability constructs are crossed with three content constructs, yielding twelve cells. On a higher level of aggregation, general intelligence integrates the primary factors for each facet.

For example, Wittmann and Süß, 1999 demonstrated that WMC has incremental validity in predicting CPS performance beyond intelligence. Bühner et al. (2008) could not confirm this result, but their study relied upon narrow operationalizations.

The second goal of the two studies presented in this paper was to investigate the relation between knowledge and complex problem solving performance. We attempted to measure knowledge about complex systems in several categories. We focused on declarative knowledge in the form of both system knowledge and action knowledge because assessing declarative knowledge is straightforward. We also attempted to measure procedural knowledge, despite the fact that no evidence has ever been put forward that responses to complex problem-solving tests exclusively reflect procedural knowledge and not declarative knowledge. Based on Cattell's investment theory (Cattell, 1987), we assumed that knowledge represents invested intelligence and examined whether the predictive effect of intelligence on CPS performance is completely mediated by prior knowledge.

We applied a CRS (i.e., a microworld with a realistic semantic embedding) in the first study, whereas we used a CAS (i.e., a microworld with an artificial semantic embedding) in the second study. Hence, the importance of preexisting knowledge with regard to CPS performance should differ between the two studies.

## STUDY 1

In the first study, we used a complex real life-oriented simulation to examine the criterion validity of intelligence, particularly reasoning, and prior knowledge for control performance in a simulated shirt factory (*Tailorshop*). As we used a very comprehensive assessment of intelligence and knowledge, we

were also interested in exploring the predictive validity of additional, more specific constructs in order to investigate the influence of the Brunswik symmetry principle (Wittmann, 1988) on the relation between intelligence, knowledge and CPS performance.

## Method Participants

One hundred and thirty-seven students from 13 high schools in Berlin took part in the experimental study in 1990 (Süß et al., 1991). They had all participated in a similar study 1 year before in which they had taken prior versions of the BIS Test and the knowledge tests and had explored the *Tailorshop* system (Süß et al., 1993a,b). Their mean age was 17.6 years ( $SD = 0.67$ ), and 40.9% were female. The participants were fully informed about the study and the voluntary nature of their participation, and anonymity was guaranteed. Written informed consent was obtained from school principals and the state school board. Subjects who withdrew from the study were required to attend other school lessons. Both Berlin studies were published in German only; a full report including the longitudinal results can be found in Süß (1996). In this paper, we report the results of the second Berlin study (here labeled Study 1) to make the results available for international readers and to discuss the two studies in the light of recent developments in CPS research.

## Materials

### Problem solving

An extended version of the *Tailorshop* system (Funke, 1983; Danner et al., 2011), originally designed by D. Dörner and first used in a published study by Putz-Osterloh (1981), was

applied as a CRS (Süß and Faulhaber, 1990). Additional minor modifications were made in the system to resolve issues with the validity of the problem-solving score that had become apparent in the study conducted 1 year before (Süß et al., 1993a,b). *Tailorshop* is a computer simulation of a shirt factory. The system has 27 variables: 10 are exogenous variables that can be manipulated directly, and 17 are endogenous variables computed by the simulation. **Figure 2** provides a screenshot of the system, and **Figure 3** an overview of the variables and their interconnections.

The system was run on a personal computer. All variables were presented in a single menu, and the values of exogenous variables could be selected via a pull-down menu. After planning all decisions, the operator ran the simulation for one virtual month. A complete trial consisted of twelve simulation cycles corresponding to 1 year of management. To obtain two independent indicators of problem solving success, participants worked on two versions of *Tailorshop* with different starting values corresponding to different shirt factories and different economic conditions. Problem solving performance was measured by participants' total assets after 12 simulated months. Since the distribution of raw scores deviated considerably from a normal distribution, we transformed them into rank scores and aggregated participants' ranks from the two simulation runs into one total score.

### Intelligence test

To assess intellectual abilities, we used a prior version of the BIS Test (Jäger et al., 1997; for a full English description see Süß and Beauducel, 2015; for prior test versions see Süß, 1996). This test consists of three to five different tasks for each of the 12 cells in the matrix structure of the BIS. Each task assigned to a cell in the model is used to measure one operation ability as well as one content ability. The four operation abilities are thus measured with scales consisting of 9–15 tasks each and balanced over the three content categories. Analogously, content abilities are measured with scales consisting of 15 tasks across the four different operation abilities. Thus, the same variables are used in different ways for different scales. The scales for one facet are built by aggregating variables that are distributed in a balanced way over the other facet. This suppresses unwanted variance, i.e., the variance associated with factors from the other facet (Wittmann, 1988). However, the scores for operation abilities and content abilities are not statistically independent. An indicator of general intelligence is built by aggregating either the operation scores or content scores.

### Knowledge tests

Preexisting general economics knowledge was assessed with an age-normed economics test (Deutsche Gesellschaft für Personalwesen [DGP], 1986, with a few questions added from the economics test from Krumm and Seidel, 1970)<sup>2</sup>. The questionnaire consisted of 25 multiple-choice items on the meaning of technical terms from the domain of economics.

A new test was developed to assess system-specific knowledge about *Tailorshop* (Kersting and Süß, 1995). This test had two parts, one for system knowledge and one for action knowledge.

System knowledge refers to knowledge about features of individual variables (e.g., development over time, degree of connectedness with other variables) and about relationships between variables in a system. The system knowledge part of the test was developed in accordance with test construction principles for optimizing content validity (Klauer, 1984; Haynes et al., 1995). It consisted of three scales:

- (1) Multiple choice questions about the connections between two variables. One out of six statements in the following form had to be selected as correct:
  - (a) An increase in variable X increases variable Y.
  - (b) An increase in variable X decreases variable Y.
  - (c) An increase in variable Y increases variable X.
  - (d) An increase in variable Y decreases variable X.
  - (e) Variable X and variable Y interact, that is, they both depend on one another.
  - (f) (a) through (e) are false.

There were 20 questions of this type.

- (2) Questions about hypotheses concerning single variables: Participants had to evaluate statements about the regular behavior of individual system variables, e.g., “The price of shirts rises and falls by chance” (which is false) or “Production depends – among other factors – on my workers' motivation, which in turn depends on the level of wages” (which is true). The scale consisted of 25 independent items.
- (3) Arrow test for connections among multiple variables: Sets of four variables were represented by labeled boxes in a diamond-shaped arrangement. Participants had to draw arrows connecting the variables that had a direct causal connection in the system, and designate the direction of correlation with a plus or minus sign (as in **Figure 3**). Each of the six possible pairings in a set was counted as an independent item that was marked as either correct or incorrect, yielding a total of 42 items.

Action knowledge refers to knowledge about appropriate actions in a certain situation, given a certain goal. It was assessed in this study via two subtests. The test of declarative action knowledge presented “rules of thumb” for successfully managing the *Tailorshop* simulation, which had to be evaluated as correct or incorrect. Half of the 12 rules were correct, i.e., they were helpful in obtaining high total assets within 12 months, while the other half were incorrect.

In the second subtest, participants were given a system state in the form of a screen display. They were given the goal of maximizing or minimizing a certain system variable, for example, minimizing the number of shirts in the store. They had to select which one out of six alternative decision patterns would be best-suited to reaching this goal in the next simulation cycle. This subtest consisted of six items with different system states, goals, and decision options. In contrast to the

<sup>2</sup>Participants only took the economics test in the first Berlin study, i.e., these data were assessed 1 year before all others reported here.



Hier der Zustand ihres Unternehmens am Ende von Monat 1

	Planung	aktuell	Vormonat		aktuell	Vormonat
Rohmaterial		802↓	1046	Gesamtvermögen	246868↓	249457
Hemdpreis DM		80↑	76	Kontostand	182797↑	182377
Werbung DM		2900↑	2800	Rohmaterialpreis DM	2.06↓	5.23
Verkaufsstellen		1	1	Nachfrage	195↓	215
Handelsvertreter		2	2	Hemden am Lager	176↑	127
50er-Maschinen		10	10	Verkaufte Hemden	195↓	215
100er-Maschinen		0	0	Produktion	244↓	261
Wartung DM		1100↓	1200	Produktionsausfall %	86↑	85
Arbeiter an 50er		8	8	Maschinenschäden %	12↑	6
Arbeiter an 100er		0	0	Maschinenauslastung %	49↓	52
Lohn DM		1095↑	1080	Arbeitsauslastung %	61↓	65
Sozialkosten DM		45↓	50	Arbeitsmotivation %	30↑	28

1 Sie können

↓ Rohmaterial für Hemden einkaufen

E

1 Mögliche Aktionen

←J Menüpunkt aufrufen

↓ nächster Menüpunkt

↑ voriger Menüpunkt

v Vormonat ausblenden

a Planung ausführen

E

Restzeit: 19 Minuten

FIGURE 2 | Screenshot of the exploration phase of the Tailorshop system as applied in Study 1.

declarative questions, this task did not require participants to explicitly declare rules for action. Instead, the rules governing their decision-making remained implicit, providing a good opportunity to capture task-relevant procedural knowledge. Thus, we will refer to this subscale as procedural action knowledge.

Sum scores were built for each subtest and a total score was calculated by aggregating the subtest scores, weighted equally.

Each type of question was introduced by the experimenter with one or two examples. There was no time limit, but participants were instructed not to spend too much time on any single question.

## Procedure

The students took tests on 2 days for 5–6 h each. On the first day, they worked on the BIS Test and the general economics test as well as some further questionnaires. Testing was done in groups of 20–30 in school classrooms. On the second day, participants were first introduced to the *Tailorshop* system via detailed instructions, including two standardized practice cycles guided by the experimenter. Afterward, the students in the sample were randomly divided into three groups, and two groups were given additional opportunities to acquire system-specific knowledge.<sup>3</sup> Next, system-specific knowledge was assessed (time T1) by instructing participants to build hypotheses about *Tailorshop* on basis of their (superficial)

experience with the system. Participants then tried to manage the *Tailorshop* twice for 12 simulated months. Finally, system-specific knowledge was tested again (time T2). The knowledge test took about 80 min the first time and about 60 min the second time. Each problem-solving trial lasted about 50 min. The participants took these tests in smaller groups at the university's computer lab.

## Results

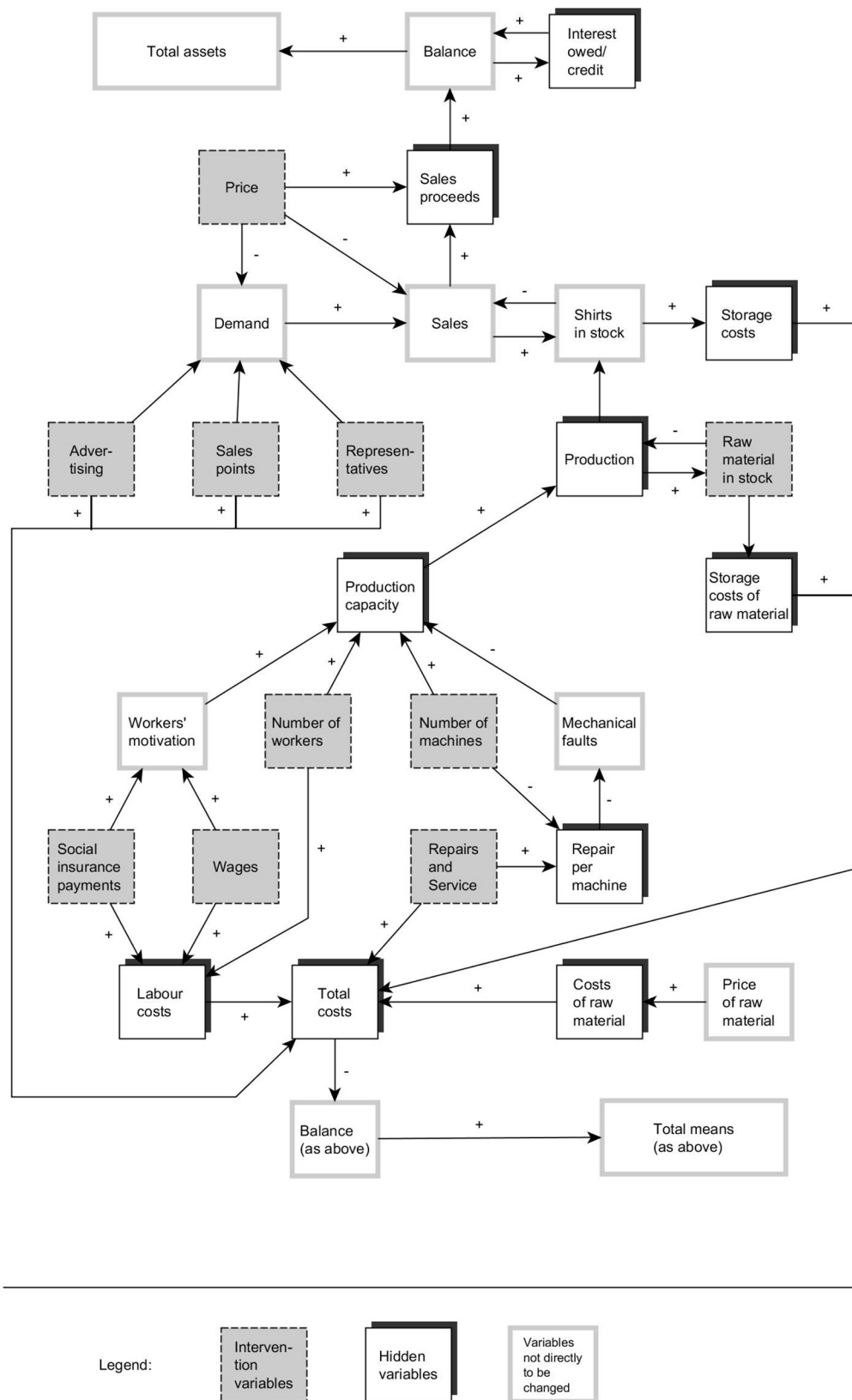
We will first present the results of separate analyses of the relationship between problem-solving performance and different groups of predictors. Then, we integrate all the variables into a path model. Ten participants had missing data for the economics knowledge test. Thus, we applied the full information maximum likelihood (FIML) procedure to account for the missing data. See Table 1 for descriptive statistics and the full correlation matrix.

## Complex Problem Solving and Intelligence

The parallel-test reliability of problem-solving performance was  $r = 0.67$  ( $p < 0.01$ ). This indicates that the criterion measures had satisfactory reliability and justifies their aggregation into a single score. Two multivariate regressions were computed with the aggregated performance criterion, first with the four operation scales and then with the three content scales of the BIS as predictors. The results are summarized in Table 2 (upper half, correlations in brackets).

Among the operation scales, reasoning ( $r = 0.34$ ,  $p < 0.01$ ) was as expected significantly correlated with problem-solving success, furthermore, creativity ( $r = 0.22$ ,  $p = 0.01$ ) as well. In the regression model, however, only reasoning had a significant beta weight ( $\beta = 0.43$ ,  $p < 0.01$ ). Among the content scales, only

<sup>3</sup>The first group could explore the system for 30 min on their own (exploration group), while the second group could study the system's causal model for 30 min following standardized instructions (instructions group). The third group had no opportunity to acquire additional system-specific knowledge (control group). In this paper, we use the results for the full sample without considering the experimental variations. Experimental results and group-specific results are reported in Süß (1996).



**FIGURE 3 |** The causal structure of the Tailorshop system.

**TABLE 1 |** Study 1: Means, standard deviations, and correlations.

Variable	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7	8	9	10
(1) BIS: g	−0.03	6.76										
(2) BIS: Speed	−0.02	2.30	0.77**									
(3) BIS: Memory	−0.02	2.36	0.66**	0.31**								
(4) BIS: Creativity	0.00	2.22	0.66**	0.38**	0.21*							
(5) BIS: Reasoning	0.00	2.51	0.79**	0.53**	0.35**	0.35**						
(6) BIS: Verbal	0.00	2.60	0.78**	0.57**	0.52**	0.57**	0.59**					
(7) BIS: Figural	−0.03	2.55	0.83**	0.62**	0.52**	0.61**	0.64**	0.52**				
(8) BIS: Numerical	−0.00	3.12	0.84**	0.69**	0.56**	0.46**	0.69**	0.44**	0.55**			
(9) Know: General	0.02	1.77	0.10	−0.12	−0.01	0.20*	0.21*	0.11	0.14	0.01		
(10) Know: Dec. Sys. t1	1.81	0.44	0.24**	0.08	0.06	0.07	0.46**	0.18*	0.18*	0.23**	0.25**	
(11) Know: Dec. Act. t1	5.69	1.73	−0.00	0.04	−0.18*	0.07	0.06	−0.08	0.01	0.05	0.21*	0.11
(12) Know: Pro. Act. t1	11.15	3.47	0.10	0.07	0.08	0.00	0.13	0.04	0.04	0.16	0.12	0.15
(13) Know: Spec. Tot. t1	72.75	13.07	0.25**	0.10	0.04	0.10	0.46**	0.18*	0.18*	0.24**	0.30**	0.93**
(14) Know: Dec. Sys. t2	1.87	0.40	0.30**	0.13	0.13	0.06	0.53**	0.21*	0.21*	0.31**	0.25**	0.83**
(15) Know: Dec. Act. t2	6.94	1.82	0.16	0.15	−0.01	0.03	0.29**	0.06	0.15	0.18*	0.15	0.27**
(16) Know: Pro. Act. t2	11.83	3.41	0.07	0.06	0.08	−0.07	0.13	−0.05	0.08	0.13	0.14	0.17*
(17) Know: Spec. Tot. t2	76.84	12.37	0.32**	0.18*	0.12	0.05	0.55**	0.19*	0.24**	0.35**	0.28**	0.80**
(18) CPS	138.00	72.58	0.22*	0.08	−0.03	0.22*	0.34**	0.11	0.16	0.25**	0.36**	0.43**
(19) Gender	1.41	0.49	−0.04	0.04	0.08	−0.05	−0.17	0.15	−0.04	−0.17	−0.42**	−0.32**
(20) Age	17.55	0.67	−0.21*	−0.20*	−0.12	−0.08	−0.19*	−0.18*	−0.16	−0.17	0.09	−0.25**

**TABLE 1 |** Continued

Variable	<i>M</i>	<i>SD</i>	11	12	13	14	15	16	17	18	19
(11) Know: Dec. Act. t1	5.69	1.73									
(12) Know: Pro. Act. t1	11.15	3.47	0.16								
(13) Know: Spec. Tot. t1	72.75	13.07	0.28**	0.41**							
(14) Know: Dec. Sys. t2	1.87	0.40	0.07	0.07	0.75**						
(15) Know: Dec. Act. t2	6.94	1.82	0.49**	0.18*	0.34**	0.24**					
(16) Know: Pro. Act. t2	11.83	3.41	0.17*	0.54**	0.30**	0.17*	0.14				
(17) Know: Spec. Tot. t2	76.84	12.37	0.18*	0.22*	0.79**	0.94**	0.38**	0.41**			
(18) CPS	138.00	72.58	0.36**	0.24**	0.51**	0.37**	0.28**	0.29**	0.46**		
(19) Gender	1.41	0.49	−0.15	−0.06	−0.33**	−0.38**	−0.14	−0.11	−0.38**	−0.35**	
(20) Age	17.55	0.67	−0.06	0.05	−0.21*	−0.19*	−0.22**	−0.02	−0.19*	−0.24**	−0.02

\* Indicates  $p < 0.05$ ; \*\* indicates  $p < 0.01$ . *M* and *SD* are used to represent mean and standard deviation, respectively. BIS, Berlin Intelligence Structure Test; Know: General, general knowledge (economics); Know: Dec. Sys, declarative system knowledge; Know: Dec. Act., declarative action knowledge; Know: Pro. Act., procedural action knowledge; Know: Spec. Tot., total problem-specific knowledge; CPS, complex problem solving (Tailorshop); t1, measurement at the Time 1; t2, measurement at the Time 2.

**TABLE 2 |** Multiple regression of problem solving performance on the operation, content, and total scales of the BIS.

	Speed	Mem.	Creat.	Reas.	$R^2_{adj}$	Verb.	Fig.	Num.	$R^2_{adj}$	g	$R^2$
Study 1: Tailorshop	−0.16 (0.08)	−0.16 (−0.03)	0.16 (0.22*)	0.43* (0.34*)	0.15*	0.04 (0.16)	−0.01 (0.11)	0.22* (0.25*)	0.04*	0.22 (0.22)	0.05
Study 2: FSYS	0.02 (0.19*)	—	—	0.33* (0.34*)	0.10*	−0.18 (0.07)	0.38* (0.37*)	0.17 (0.27*)	0.16*	0.33* (0.33*)	0.10*

Beta weights in the first line; bivariate Pearson correlations in brackets in the second line. Speed, perceptual speed; Mem., memory; Creat., creativity; Reas., reasoning; Verb., verbal intelligence; Fig., figural intelligence; Num, numerical intelligence. Values with \* are significant at the 5% level.

numerical intelligence had a significant beta weight ( $\beta = 0.22$ ,  $p = 0.03$ ). The proportion of variance accounted for by the operation scales was much higher than that accounted for by the content scales, despite the fact that the two groups of predictors consisted of the same items that had merely been aggregated in different ways. Building an overall aggregate for all BIS scales

(BIS-g) only accounted for five percent of the criterion variance ( $r = 0.22$ ,  $p = 0.01$ )<sup>4</sup>, compared to 15 percent with the four

<sup>4</sup>The correlation with CPS was slightly higher ( $r = 0.27$ ) for a conventional g-score based on the factor scores of the first unrotated factor (Jensen and Wang, 1994), i.e., 7.3% of CPS variance was explained.

operation scales. In line with the Brunswik symmetry principle (Wittmann, 1988; Wittmann and Süß, 1999), this comparison shows the benefit of differentiating intellectual abilities into multiple components using a multi-faceted model. Taking the cell level of the BIS<sup>5</sup> into account, numerical reasoning was the best and thus likely the most symmetrical predictor of *Tailorshop* performance ( $r = 0.36$ ,  $p < 0.01$ ).<sup>6</sup> While the correlation between the numerical reasoning cell and the criterion was nearly the same as the correlation for reasoning, numerical reasoning was the better predictor given the substantially lower reliability of the cell score for numerical reasoning (Cronbach's  $\alpha = 0.77$ ) compared to reasoning (1-year stability,  $r = 0.90$ ,  $p < 0.01$ ). Corrected for unreliability, the true correlation was  $r = 0.43$ . In summary, aggregating repeated measures increases the reliability and thus also the validity of the CPS performance score. However, the correlations are lower than for minimally complex tasks even on the most symmetrical level ( $r = 0.58$ ), as reported in Stadler et al.'s (2015) meta-analysis.

### Complex Problem Solving and Knowledge

Four scales representing prior knowledge (time T1) were used as predictors of problem solving success in the regression analysis. These were the general economics test and the three categories of knowledge represented in the system-specific knowledge test: declarative system knowledge (measured with three subtests), declarative action knowledge (measured with the rules of thumb), and procedural action knowledge (measured using the system-states task). General economics knowledge ( $\beta = 0.21$ ,  $p < 0.01$ ;  $r_{\text{zero-order}} = 0.36$ ,  $p < 0.01$ ), declarative system knowledge ( $\beta = 0.33$ ,  $p < 0.01$ ;  $r_{\text{zero-order}} = 0.43$ ,  $p < 0.01$ ), and declarative action knowledge ( $\beta = 0.26$ ,  $p < 0.01$ ;  $r_{\text{zero-order}} = 0.36$ ,  $p < 0.01$ ) were significantly associated with problem solving performance, whereas procedural action knowledge was not ( $\beta = 0.13$ ,  $p = 0.07$ ;  $r_{\text{zero-order}} = 0.24$ ,  $p < 0.01$ ). The latter might be in part due to the low reliability of the test, which consisted of only six items. Together, general and system-specific knowledge accounted for 34 percent of the variance in CPS performance.

A significant increase in domain-specific knowledge from pre- to post-test was observed for every subscale. The strongest effect was for declarative action knowledge ( $t = 8.16$ ,  $p < 0.01$ ,  $d = 0.70$ ), with smaller effects observed for declarative system knowledge ( $t = 2.86$ ,  $p < 0.01$ ,  $d = 0.25$ ) and procedural action knowledge ( $t = 2.33$ ,  $p < 0.05$ ,  $d = 0.20$ ). Pre-post correlations were 0.83 ( $p < 0.01$ ) for declarative system knowledge, 0.49 ( $p < 0.01$ ) for declarative action knowledge, and 0.54 ( $p < 0.01$ ) for procedural action knowledge.

### An Integrative Path Model

In a second step, we tested our theoretical model via path analysis. Reasoning and general economics knowledge were assumed to be correlated exogenous variables influencing the generation of hypotheses and the acquisition of system-specific

knowledge during instruction and exploration, and thus also the amount of system-specific (prior) knowledge measured at time T1. We also assumed direct paths from reasoning, general economics knowledge and system-specific prior knowledge (T1) to control performance, and tested whether reasoning, domain-specific prior knowledge (T1) and problem-solving performance influence system-specific knowledge measured after controlling the system (T2). The resulting model is presented in Figure 4.

The path model reflects and extends the results above. System-specific prior knowledge (T1) was significantly influenced by the two correlated exogenous variables, indicating the importance of general domain knowledge, and especially of reasoning, for generating and testing hypotheses in the *Tailorshop* simulation. System-specific prior knowledge (T1) was influenced by learning processes during the instructions and, for a part of the sample, during system exploration. A total of 25.4% of the variance was explained by the two exogenous variables. General economics knowledge ( $\beta = 0.22$ ,  $p < 0.01$ ) and system-specific prior knowledge (T1;  $\beta = 0.40$ ,  $p < 0.01$ ) also had direct effects on control performance. Reasoning ability, meanwhile, had no direct effect ( $\beta = 0.12$ ,  $p = 0.12$ ), but a strong indirect effect on problem solving performance as mediated by prior knowledge. The total amount of explained variance in problem solving performance was 32%. Finally, system-specific knowledge after controlling the system (T2) primarily depended on system-specific prior knowledge (T1;  $\beta = 0.65$ ,  $p < 0.01$ ) as well as reasoning ( $\beta = 0.25$ ,  $p < 0.01$ ). Remarkably, while control performance and acquired system knowledge (T2) were substantially correlated ( $r = 0.46$ ,  $p < 0.01$ ), the direct path from control performance to acquired system-specific knowledge (T2) was not significant ( $\beta = 0.05$ ,  $p = 0.35$ ). Overall, 68.6% of the variance was explained.

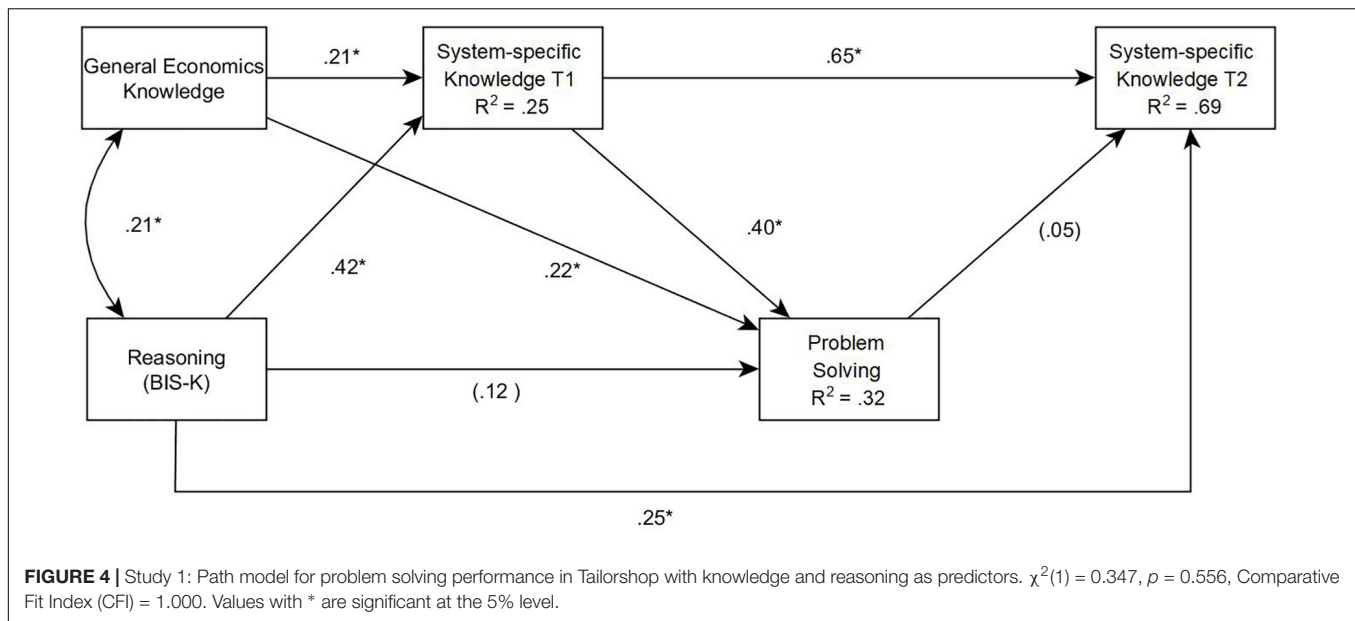
### Discussion

Both intelligence and prior knowledge were shown to be important predictors of performance controlling a complex system. Some qualifications, however, must be made to this conclusion. First, it is not general intelligence that has predictive power for problem solving success in *Tailorshop*; instead and as expected, it is the primary factor reasoning, and more specifically numerical reasoning. This underscores the importance of finding the right level of symmetry between predictor and criterion in order to estimate their true relationship (Wittmann, 1988). Second, the correlation between reasoning and problem solving performance was mediated through prior knowledge; reasoning had no direct influence on problem solving performance. This finding is in line with the results of the meta-analysis by Schmidt et al. (1986; Schmidt, 1992), which showed that the relationship between intelligence and job performance is nearly completely mediated by task-related knowledge. This may indicate that persons with higher reasoning ability have used their ability to accumulate more domain knowledge in the past. The strong relationship between reasoning and general economics knowledge supports this account. An alternative explanation is that high reasoning ability helps people transfer their general domain knowledge to the specific situation, i.e., by deriving good hypotheses about the unknown system from their general

<sup>5</sup>According to the BIS, numerical reasoning is not a more specific ability but a performance based on reasoning and numerical intelligence (Jäger, 1982).

<sup>6</sup>The correlation of CPS performance with figural reasoning was 0.26, and 0.24 with verbal reasoning.





theoretical knowledge about the corresponding domain. System-specific knowledge measured after controlling the system (T2) depends primarily on prior knowledge and reasoning. Therefore, controlling a complex system can be described as a knowledge acquisition process, providing evidence for Cattell's investment theory (Cattell, 1987). Assuming that the system has ecological validity, this finding also indicates that system-specific knowledge measured after controlling a complex system is a powerful predictor of external criteria.

The study was limited to the computer-simulated system *Tailorshop*, a microworld mainly developed by psychologists. The scenario is realistic in that it captures many psychologically relevant features of complex real-life problems, but its ecological validity as a model for a real business environment is limited. For example, real company executives spend more than 80% of their time communicating orally (e.g., Mintzberg, 1973; Kotter, 1982), a demand which was not implemented in the simulation (see Süß, 1996).

A final but important qualification to the study's results concerns reasoning in the context of knowledge. System-specific knowledge was consistently the best single predictor of problem solving success in *Tailorshop*, while general domain knowledge in economics significantly predicted additional variance. System-specific knowledge was made up of two independent predictors, declarative system knowledge and declarative action knowledge. Our study found no evidence of the dissociation between verbalized knowledge and control performance repeatedly reported by Broadbent and colleagues (Broadbent et al., 1986; Berry and Broadbent, 1988; see Berry and Dienes, 1993). *Tailorshop* is a more complex and realistic system than those used by Broadbent and colleagues. Both factors might have strongly motivated people to make use of their preexisting knowledge, i.e., to formulate explicit hypotheses for controlling the system rather than following a trial-and-error approach that would result in the acquisition of implicit knowledge.

## STUDY 2

The aim of the second study was to replicate and extend the findings presented so far. Study 2 differed from Study 1 in two important ways. First, we used the artificial world simulation *FSYS* (Wagener, 2001), which simulated a forestry company. Although *FSYS* has a rich semantic embedding and all the characteristics of complex problems, *FSYS* was developed with the aim of reducing the impact of previous knowledge of the simulated domain (i.e., general forestry knowledge) on problem solving performance. Therefore, *FSYS* can be classified as a CAS. Second, we included WMC as a further predictor. WMC is a more basic construct than reasoning and whether it is a better (i.e., more symmetrical) predictor of CPS performance than reasoning is an open question (see Zech et al., 2017). Thus, we were interested in whether one of the two constructs had incremental validity in predicting CPS performance beyond the other construct.

## Method

### Participants

One hundred fifty-nine students from the University of Magdeburg participated in the second study, which was originally conducted to evaluate a complex problem solving training (for details, see Kretzschmar and Süß, 2015), in 2010/2011.<sup>7</sup> In the present analyses, we used the full sample but excluded all non-native German speakers ( $n = 7$ ) due to the high language requirements of the intelligence test. The mean age was 23.99 years ( $SD = 4.43$ ), and 50% were female. Participants received course credit for their participation or took part in a book raffle. Participants were informed about the content of the

<sup>7</sup>A subsample was used in Kretzschmar and Süß (2015) to evaluate a CPS training. However, none of the relations between CPS and the variables used in the present study have been previously examined (for details, see the data transparency table at <https://osf.io/n2jvy>). Therefore, all analyses and findings presented here are novel.

study, the voluntary nature of participation and their ability to withdraw at any point, and that anonymity was guaranteed. All subjects provided informed consent.

## Materials

### Problem solving

We used version 2.0 of the microworld *FSYS* (Wagener, 2001). *FSYS* was developed on the basis of Dörner et al.'s (1983) theoretical framework for complex problem solving (Dörner, 1986). It is a microworld with 85 variables connected via linear, exponential, or logistic relations. The goal was to manage five independent forests in order to increase the company's value (i.e., planting and felling trees, fertilizing, pest control, etc.). Participants were first given an introduction to the program and had an opportunity to explore the system. They then managed the forest company for 50 simulated months. We used the company's total capital (i.e., an aggregated score of the five independent forests) at the end of the simulation as the performance indicator (SKAPKOR; see Wagener, 2001). Although *FSYS* simulates a forestry enterprise, the impact of prior knowledge was reduced by using abstract names for tree species, pests, fertilizer etc., and providing essential information about the artificial foresting world via an integrated information system. Previous studies have shown that *FSYS* has incremental predictive validity beyond general intelligence with regard to occupational (Wagener and Wittmann, 2002) and educational (Stadler et al., 2016) performance indicators. **Figure 5** provides a screenshot of *FSYS*.

### Intelligence

A short version of the BIS Test was used to assess intellectual abilities (Jäger et al., 1997). We specifically focused on reasoning and perceptual speed. Nine tasks were applied for each operation, balanced over the three content areas (i.e., figural, verbal, numerical; see **Figure 1**). These 18 tasks were administered according to the test manual. As in Study 1, the tasks were aggregated in order to build scales for each operation (i.e., reasoning, perceptual speed) or content (i.e., figural intelligence, verbal intelligence, numerical intelligence). An indicator for general intelligence was built by aggregating the 18 tasks in a balanced way, as described in the test handbook. Please note that the reliability of the two operative scales was lower than in Study 1; the construct validity of the three content scales and the measure of general intelligence were also reduced because no memory or creativity tasks were used. This limits the interpretability of the BIS content scales and the comparability of the results of the two studies.

### Working memory

Working memory capacity was assessed with three tasks from the computerized test battery by Oberauer et al. (2003). The *numerical memory updating* (adaptive) and *reading span* (non-adaptive) tasks measured the simultaneous storage and processing functions of working memory, whereas the *dot span* task (also named *spatial coordination*; adaptive) primarily measured the coordination function. Moreover, each content

category (i.e., figural, verbal, numerical) was represented by one task. A global score for WMC was calculated by aggregating the three equally weighted total task scores.

### Knowledge

A questionnaire to assess general forestry knowledge as a measure of preexisting domain knowledge was developed for the purpose of this study<sup>8</sup>. It covered forestry knowledge in the subdomains of tree species, soils, nutrients, damage to a forest, and silviculture. An example question was: "Which tree is not a conifer?" The 22 multiple-choice items were scored dichotomously. Four items were excluded due to poor psychometric properties (i.e., a low item-total correlation). The remaining 18 items were aggregated to form a global sum score.

To assess system-specific knowledge about *FSYS*, we used Wagener's (2001) knowledge test about the microworld. The 11 multiple-choice items addressed system and action knowledge across all relevant areas of *FSYS*. For example: "A forest is infested by vermin XY. Which procedure would you apply?" In order to limit the number of questions, we did not differentiate between different types of knowledge. Therefore, we used a sum score as a global indicator of system-specific knowledge.

### Procedure

Participants took part in two sessions each lasting about 2.5 h. All testing was done in groups of up to 20 persons at the university computer lab. The first session comprised tests of intelligence and WMC. In the second session, participants completed tests of general forestry knowledge, complex problem solving, and system-specific knowledge. In contrast to Study 1, system-specific knowledge was assessed only once, after participants had worked with the CPS scenario (similar to Wagener, 2001). As the study was originally designed as an experimental training study (see Kretzschmar and Süß, 2015), the procedure differed slightly between the two experimental groups. About half of the participants completed the second session the day after the first session. The other half participated in a CPS training in between and completed the second session about 1 week after the first session.

## Results

We will first present results for individual groups of predictors of CPS performance before integrating the results into a combined path model. Due to the original study design (i.e., exclusion criteria for the training, dropout from the first session to the second), up to 24% of the data for the knowledge tests and the CPS scenario were missing. We used the full information maximum likelihood (FIML) procedure to account for missing data. The smallest sample size in the analyses of individual groups of predictors was 116. The data are publicly available via the Open Science Framework<sup>9</sup>. See **Table 3** for descriptive statistics and the full correlation matrix.

<sup>8</sup>We would like to thank Clemens Leutner for professional advice in developing the questionnaire.

<sup>9</sup><https://osf.io/n2jvy>



**FIGURE 5 |** Screenshot of the exploration phase of FSYS system as applied in Study 2.

**TABLE 3 |** Study 2: Means, standard deviations, and correlations.

Variable	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7	8	9	10	11
(1) BIS: g	0.04	0.50											
(2) BIS: Speed	0.04	0.60	0.82**										
(3) BIS: Reasoning	0.04	0.57	0.88**	0.47**									
(4) BIS: Verbal	0.08	0.60	0.71**	0.69**	0.60**								
(5) BIS: Figural	0.02	0.62	0.84**	0.69**	0.71**	0.47**							
(6) BIS: Numerical	0.01	0.69	0.78**	0.67**	0.68**	0.35**	0.46**						
(7) WMC	0.12	2.10	0.55**	0.41**	0.52**	0.33**	0.45**	0.48**					
(8) Know: General	8.35	2.74	0.08	0.00	0.13	0.04	0.03	0.09	0.02				
(9) Know: Specific	5.32	1.94	0.36**	0.17	0.41**	0.20*	0.35**	0.23*	0.22*	0.19*			
(10) CPS	57.59	22.51	0.33**	0.19*	0.34**	0.07	0.37**	0.27**	0.32**	0.16	0.51**		
(11) Age	23.99	4.43	−0.34**	−0.21**	−0.34**	−0.23**	−0.36**	−0.17*	−0.31**	0.24*	−0.17	−0.21*	
(12) Gender	0.50	0.50	−0.17*	−0.05	−0.20*	0.13	−0.09	−0.35**	−0.06	−0.18	−0.20*	−0.10	−0.08

\* Indicates  $p < 0.05$ ; \*\* indicates  $p < 0.01$ . *M* and *SD* are used to represent mean and standard deviation, respectively. BIS, Berlin Intelligence Structure Test; WMC, working memory capacity; Know: General, general forestry knowledge; Know: Specific, system-specific knowledge; CPS, complex problem solving performance (FSYS).

## Complex Problem Solving, Intelligence, and Working Memory

The results of two multivariate regressions of FSYS performance scores on the BIS operative and content scales, respectively, are summarized in **Table 2** (lower half, correlations in brackets).

The results for operation abilities are similar to those from the first study, with reasoning the only significant predictor ( $\beta = 0.33$ ,  $p < 0.01$ ). However, figural intelligence was the only statistically significant predictor among the content scales ( $\beta = 0.38$ ,  $p < 0.01$ ). This seems plausible given that

FSYS displays important information graphically rather than numerically (e.g., diagrams showing the forestry company's development). However, a large amount of information is also presented numerically, meaning that numerical reasoning should exert an influence as well. Taking the cell level of the BIS into consideration: Numerical reasoning (Cronbach's  $\alpha = 0.66$ ) became similarly strongly associated with FSYS control performance ( $r = 0.37$ ,  $p < 0.01$ ; corrected for unreliability  $r = 0.46$ ) as figural reasoning (Cronbach's  $\alpha = 0.72$ ;  $r = 0.36$ ,  $p < 0.01$ ; corrected for unreliability  $r = 0.42$ ). Verbal reasoning (Cronbach's  $\alpha = 0.51$ ) remained unassociated with FSYS performance ( $r = 0.02$ ,  $p = 0.82$ ). In contrast to Study 1, the content scales accounted for a slightly larger share of the variance in FSYS (16%) than the operation scales (10%). General intelligence (BIS-g) had a  $0.33$  ( $p < 0.01$ ) correlation with problem solving performance.

Next, we compared the impact of reasoning and WMC as predictors of success in FSYS. Both predictors exhibited an almost equal and statistically significant zero-order correlation ( $r_{\text{BIS-R.FSYS}} = 0.34$ ,  $p < 0.01$ ;  $r_{\text{WMC.FSYS}} = 0.32$ ,  $p < 0.01$ ). In hierarchical regressions, each explained a similar but non-significant amount of incremental variance over and above the other predictor ( $\Delta R^2_{\text{BIS,K}} = 0.02$ ;  $\Delta R^2_{\text{WMC}} = 0.02$ ). The total explained variance was 12.2% (adjusted). In summary, working memory did not increase the statistical significance of the multiple correlation when entered as a second predictor.

### Complex Problem Solving and Knowledge

General forestry knowledge was not significantly correlated with FSYS performance ( $r = 0.16$ ,  $p = 0.09$ ). Thus, the (non-)impact of prior domain knowledge in FSYS was similar as in previous studies ( $r = 0.13$ ; Wagener, 2001), emphasizing how the impact of prior knowledge depends on the specific type of microworld (i.e., CRS in Study 1 vs. CAS in Study 2). The correlation between system-specific knowledge (measured after working on FSYS) and FSYS performance was  $r = 0.51$  ( $p < 0.01$ ).

### An Integrative Path Model

In line with our assumptions about the relations among the predictor and criterion variables and building upon the results of the first study, we constructed a path model to integrate our findings. Perceptual speed from the BIS Test was excluded from the analyses because it was not significantly associated with any endogenous variable when controlling for reasoning. Prior general forestry knowledge was also omitted from the path model for the same reason.

In the first model (Figure 6, Model A), working memory had a direct influence on reasoning but not on FSYS control performance and system-specific knowledge. In this model [ $\chi^2(2) = 4.538$ ,  $p = 0.10$ , CFI = 0.977, SRMR = 0.038], control performance ( $\beta = 0.34$ ,  $p < 0.01$ ) and acquired system-specific knowledge about the microworld FSYS ( $\beta = 0.26$ ,  $p < 0.01$ ) were significantly influenced by reasoning. The total amount of explained variance for control performance and system-specific knowledge were 11% and 32%, respectively.

In a second (fully saturated) model (Figure 6, Model B: dashed lines and coefficients in brackets), direct paths from working

memory to FSYS control performance and system-specific knowledge were added. In this model, working memory had a small but non-significant direct effect on control performance ( $\beta = 0.20$ ,  $p = 0.09$ ), i.e., the effect of working memory is primarily based on its shared variance with reasoning. Furthermore, WMC functioned as a suppressor when it came to predicting system-specific knowledge. In other words, despite the positive zero order correlation between the two variables (see above), the direct path from WMC to system-specific knowledge was negative ( $\beta = -0.13$ ,  $p = 0.19$ ), while the impact of reasoning on system-specific knowledge slightly increased ( $\beta = 0.33$ ,  $p < 0.01$ ). On the other hand, the path from working memory to system-specific knowledge was statistically non-significant, and the explained variance in system-specific knowledge did not significantly increase [ $\Delta R^2 = 0.012$ ,  $F(1,148) = 2.663$ ,  $p = 0.46$ ].

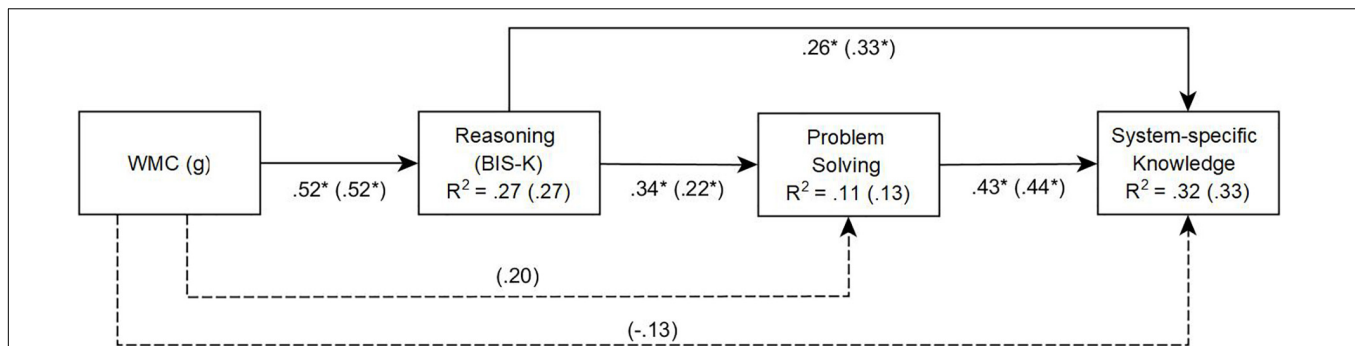
### Discussion

The general findings of Study 1 with regard to the impact of intelligence on CPS performance could be replicated in Study 2. However, as Study 2 was conducted with a different microworld with different cognitive demands (e.g., less relevance of prior knowledge), the results differed somewhat compared to those of Study 1.

With regard to intelligence, reasoning was again the strongest and sole predictor of CPS performance. Because general intelligence (g) was operationalized substantially more narrowly than in Study 1, the results for reasoning and g were comparable. These findings highlight the effect of the specific operationalization of intelligence selected. If intelligence is broadly operationalized, as proposed in the BIS (see Study 1), the general intelligence factor is not equivalent to reasoning (aka fluid intelligence; see also Carroll, 1993; McGrew, 2005; Horn, 2008) and different results for g and for reasoning in predicting CPS performance can be expected (see e.g., Süß, 1996). With regard to the content facet, FSYS shared the most variance with figural intelligence. However, the cell level of the BIS provided a more fine-grained picture: figural reasoning was just as highly correlated with FSYS performance as numerical reasoning. Although Study 1 and Study 2 must be compared with caution (i.e., due to different operationalizations of the BIS scales, see Figure 1, and limited BIS reliability on the cell level), it is clear that different CPS tests demand different cognitive abilities. At the same time, these findings highlight the importance of the Brunswik symmetry principle (Wittmann, 1988; Wittmann and Süß, 1999). A mismatch between predictor and criterion (e.g., figural reasoning and *Tailorshop* performance in Study 1; or numerical intelligence and FSYS performance in Study 2) substantially reduces the observed correlation (for another empirical demonstration in the context of CPS, see Kretzschmar et al., 2017). Ensuring that the operationalizations of the constructs are correctly matched provides an unbiased picture of the association across studies (Zech et al., 2017).

Working memory capacity was strongly related to reasoning and largely accounted for the same portion of variance in problem solving success as reasoning; it did not explain substantial variance over and above reasoning. These results complement the mixed pattern of previous findings, in which





**FIGURE 6 |** Study 2: Path Model A for problem solving performance and system-specific knowledge in FSYS, predicted by reasoning and working memory capacity (WMC). Fit for model A (without dashed lines):  $\chi^2(2) = 4.538$ ,  $p = 0.10$ , Comparative Fit Index (CFI) = 0.977. Path model B (saturated) with dashed lines and values in brackets. Values with \* are significant at the 5% level.

working memory explained CPS variance above and beyond intelligence (Wittmann and Süß, 1999), was the only predictor of CPS variance when simultaneously considering figural reasoning (Bühner et al., 2008), but did not explain CPS variance above and beyond reasoning (Greiff et al., 2016). In our view, there is little unique criterion variance to explain because the predictors are highly correlated. Even small differences in operationalization or random fluctuations can make one or the other predictor dominate (for a different view, see Zech et al., 2017).

Preexisting knowledge (i.e., general forestry knowledge) did not contribute to problem solving success. This finding highlights the importance of the CPS measurement approach selected. Whereas *Tailorshop* was developed as a complex real life-oriented simulation in which prior domain knowledge plays a substantial role, FSYS was developed with the aim of reducing the influence of prior knowledge (Wagener, 2001). Therefore, in addition to the distinction between microworlds and MCS, the differential impact of prior knowledge in terms of semantic embedding has to be considered when examining the validity of CPS (e.g., the effects might differ for CRS vs. CAS, as in the present study). It should be noted that in Stadler et al.'s (2015) meta-analysis, a study featuring FSYS (in which prior knowledge has no impact) and a study involving a virtual chemistry laboratory (in which prior knowledge has an effect; see Scherer and Tiemann, 2014) were both classified as single complex system studies. As a substantial portion of the variance in CPS performance in semantically embedded microworlds can be attributed to prior knowledge, the question arises as to whether a more fine-grained classification of the CPS measures in Stadler et al.'s (2015) meta-analysis would have resulted in different findings. In summary, the heterogeneity of different CPS measurements makes it difficult to compare studies or conduct meta-analyses (some would say impossible, see Kluwe et al., 1991).

## GENERAL DISCUSSION

The presented studies had two main goals. First, we wanted to investigate the predictive validity of differentiated cognitive

constructs for control performance in complex systems. Second, we were interested in how preexisting general knowledge and system-specific prior knowledge contribute to successful system control.

Both studies clearly demonstrate that intelligence plays an important role in control performance in complex systems. This is in contrast to former claims in early CPS research that problem solving success in complex, dynamic, partially intransparent systems is not at all correlated with intelligence test scores (e.g., Kluwe et al., 1991). Our results point to several explanations for prior failures to find positive correlations. First, previous studies used only a single problem solving trial, meaning that the performance criterion presumably was not satisfactorily reliable. Second, several previous studies did not differentiate between different aspects of intelligence, but used a measure of general intelligence. In our studies, however, general intelligence (g) as conceptualized in the BIS and operationalized with the BIS Test was not a good predictor of control performance. Instead and as was expected, the second-order construct of reasoning, and more specifically numerical reasoning, had the strongest relationship with success in the complex real-world oriented system (*Tailorshop*), while figural and numerical reasoning had the strongest relationship with success in the complex artificial world problem (FSYS). However, whether g and reasoning are distinguishable from each other (Carroll, 1993), and thus also whether the two differ in predicting CPS performance, depends on the level of generality, i.e., the broadness of the operationalization of g.

Our results are in line with the first Berlin study (Süß et al., 1993a,b) and several other studies using the *Tailorshop* system and other CRSs focusing on ecological validity (e.g., Wittmann and Süß, 1999; Kersting, 2001; Leutner, 2002; Rigas et al., 2002; Ryan, 2006; Danner et al., 2011), and were confirmed in Stadler et al.'s (2015) meta-analysis.

## Is There Evidence for a New Construct 'Complex Problem Solving Ability'?

The two presented studies, however, are limited to one microworld each, and do not answer broader questions

regarding generalizability. In particular, the convergent validity of microworlds was not addressed, but this question is essential for postulating *complex problem solving ability* as a new ability construct.

The following criteria must be considered in justifying a new ability construct (cf., Süß, 1996, 1999): (1) temporal stability, (2) a high degree of generality (i.e., the construct can be operationalized across different tasks, showing convergent validity), (3) partial autonomy in the nomological network of established constructs (i.e., the shared performance variance in different tasks cannot be explained by well-established constructs), and (4) evidence for incremental criterion validity compared to established constructs. In this section, we briefly review the empirical results regarding the existence of a unique CPS construct. We focus on CPS research utilizing CRS (i.e., microworlds with semantic embeddings)<sup>10</sup>.

The 1-year stability of CRS performance in the Berlin study (see Süß, 1996) was  $r = 0.49$ , which is substantial, but much lower than that for the intelligence constructs. The temporal stability of the BIS scales ranged from 0.65 for creativity to 0.90 for reasoning. In addition, the time-stable performance variance was explained completely by intelligence and prior knowledge (Süß, 1996). To the best of our knowledge, no results on temporal stability for other CRS and temporal stability for aggregated scores based on different CRS are currently available.

Wittmann et al., (1996; Wittmann and Süß, 1999; Wittmann and Hattrup, 2004) investigated the convergent validity of CRS. Wittmann et al. (1996) applied three different CRS (*PowerPlant*, *Tailorshop*, and *Learn!*), the BIS Test and domain-specific knowledge tests for each system to a sample of university students. The correlations of the CRS were significant but rather small (0.22–0.38), indicating low convergent validity<sup>11</sup>. However, because the reliability of each CRS was substantially higher than their intercorrelations, substantial system-specific variance has to be assumed. Performance on each of the three systems was predicted by reasoning and domain-specific prior knowledge to a substantial degree. In a structural equation model with a nested-factor BIS model (Schmid and Leiman, 1957; Gustafsson and Balke, 1993) as predictor, the CPS g-factor with two performance indicators for each of the three systems (i.e., the CPS ability construct) was predicted by general intelligence ( $\beta = 0.54$ ), creativity (0.25) and reasoning (0.76), whereas perceptual speed and memory did not contribute to prediction (Süß, 2001)<sup>12</sup>. In this model, reasoning, though orthogonal to general intelligence,

was the strongest predictor of the complex problem solving ability factor. Almost all of the variance could be explained by the BIS, putting the autonomy of the CPS construct into question.

In sum, there is no evidence for a new ability construct based on CRSs. This, however, does not mean that this kind of research cannot provide important new insights into CPS processes (see Süß, 1999), and that CPS performance cannot predict real-life performance beyond psychometric intelligence measures to a certain extent (e.g., Kersting, 2001; Danner et al., 2011).

Kersting (2001) predicted police officers' job performance over 20 months on the basis of intelligence (short scales for reasoning and general intelligence from the BIS Test), CPS performance (two simulations, including *Tailorshop*), and acquired system-specific knowledge (measured after controlling the system). In a commonality analysis (Kerlinger and Pedhazur, 1973), 24.9% of job performance variance was explained. The strongest specific predictor was intelligence (7.3%); reasoning and general intelligence at about the same level); CPS performance and system-specific knowledge explained 3.9 and 3.0% of the overall criterion, respectively. The largest share of the variance was confounded variance between intelligence and system-specific knowledge (24.9%). In comparison to our first study, both intelligence scales had reduced predictive validity due to lower reliabilities. However, this study shows that exploring and controlling CRS must be considered a learning process. Acquired system knowledge represents invested intelligence (i.e., crystallized intelligence) and was a small but additional predictor of real-life performance beyond intelligence. This provides that ecological-valid complex systems can additionally predict external criteria, and are useful learning and training tools for acquiring domain-specific knowledge.

## PART II: REVIEW AND CRITIQUE OF THE MINIMALLY COMPLEX SYSTEM (MCS) APPROACH

The research presented and discussed in the first part of the paper focuses on CRSs. From the beginning, CRS research was criticized for numerous reasons, including the lack of a formal description of the system, the lack of an optimal solution as an evaluation criterion for subjects' behavior and performance, the uncontrolled influence of prior knowledge, low or unknown reliability of the scores, and low or even non-existent convergent validity and predictive validity with respect to relevant external criteria (for summaries, see e.g., Funke, 1995; Süß, 1996; Kluge, 2008a). Therefore, the MCS approach (Greiff et al., 2012) was developed to overcome the limitations of former microworlds. The MCS approach is remarkably prominent in recent CPS research, which may be a consequence of the higher reliability and validity such systems are assumed to have in comparison to CRS (e.g., Greiff et al., 2015b). Consequently, some might argue that research on CPS performance based on CRS, as presented in the first part of the paper, is less reliable and informative. However, whether the MCS approach is really a superior alternative to

<sup>10</sup>For a review focusing on CPS research applying the minimally complex systems (MCS) approach, see Kretzschmar and Süß (2015).

<sup>11</sup>In the study of Ryan (2006) with 298 University students the intercorrelations of three scenarios, *Furniture Factory* (FF), *Tailorshop* (T) and *FSYS* (F), were also rather small but significant ( $r_{FF,T} = 0.30$ ,  $r_{FF,F} = 0.27$ ,  $r_{T,F} = 0.10$ ; Stankov, 2017).

<sup>12</sup>The structural equation model by Süß (2001) is copied in Wittmann and Hattrup (2004) as Figure 6. This model was built in two steps: First, BIS and CPS-g were modeled separately. Specific CPS factors for the three systems were not modeled because only two indicators were available for each system. Instead, the errors of the two indicators in each system were allowed to correlated as system-specific variance. Second, the five BIS factors (g and the four operative abilities) were used to predict CPS-g. Fit statistics for the final model are not valid because the loadings of both measurement models were optimized in the first step.

studying problem solving in complex situations remains up for debate.

The MCS approach updates and further develops ideas that have been present since the beginning of CPS research. Funke (1993) suggested artificial dynamic systems as a research tool based on systems of linear equations. Buchner and Funke (1993) proposed the theory of finite state automata as a tool for developing CPS tasks. Applying this, Kröner (2001; Kröner et al., 2005), for example, implemented *MultiFlux*, which simulates a fictitious machine, within the finite-state framework. This idea was further developed into MCS, e.g., *Genetics lab* (Sonnleitner et al., 2012) and *MicroDYN* (Greiff et al., 2012). Generally, about 9–12 artificial world tasks, tiny systems with up to three exogenous and three endogenous variables each, are applied in three phases: (1) free system exploration, (2) knowledge acquisition (i.e., assessment of acquired system knowledge), and (3) knowledge application (i.e., assessment of action knowledge). The required testing time is less than 5 min for each minimal system. Each system provides three scores, one for each of the above-mentioned phases, which are then used to form three corresponding knowledge scales. According to our knowledge taxonomy, Phase 2 measures declarative system knowledge (i.e., relations between variables), while Phase 3 measures procedural action knowledge (i.e., system interventions in order to achieve a given goal). The items in these two subtests are similar to the items in the arrows task and the system-states task of the *Tailorshop* knowledge test. Whereas each item in the MCS scales refers to a different minimal system, all items in the *Tailorshop* knowledge test refer to the same system. Nevertheless, the MCS tasks are very similar to each other and implement only a small number of CPS characteristics, giving the subtests high internal consistencies. Specifically, all minimal systems can be fully explored with the simple strategy “vary one thing at a time” (VOTAT; e.g., Vollmeyer et al., 1996) or the closely related strategy “vary one or none at a time” (Beckmann and Goode, 2014; for additional distinctions see Lotz et al., 2017). No special training is necessary to learn these strategies. Instead, they can be learned by instruction or examples of correct and incorrect applications. On the other hand, these strategies are clearly not sufficient for exploring CRS, i.e., systems with many exogenous variables, indirect and side effects, delayed effects, and eigendynamics, especially if the time for the task is limited or in real-time simulations (e.g., Bremer’s fire-fighter; Rigas et al., 2002). For the latter, the quality of one’s hypotheses, which is based on domain knowledge, is a necessary prerequisite for successfully exploring the system. In summary, the features of MCS measurements outlined here, along with further criticisms of this approach (e.g., Funke, 2014; Scherer, 2015; Schoppek and Fischer, 2015; Dörner and Funke, 2017; Funke et al., 2017; Kretzschmar, 2017), substantially narrow the validity of the MCS approach as an indicator of CPS.

On the other hand, the relevance of the MCS approach is shown by many studies that have modeled the internal structure of MCS tasks (e.g., Greiff et al., 2012; Sonnleitner et al., 2012), provided evidence that performance variance cannot be sufficiently explained by reasoning (e.g., Wüstenberg et al.,

2012; Sonnleitner et al., 2013; Kretzschmar et al., 2016), found strong convergent validity as well as a lower correlation with a CRS (i.e., *Tailorshop*; Greiff et al., 2015b; for a different view, see Kretzschmar, 2017), and demonstrated incremental validity in predicting school grades beyond reasoning (e.g., Greiff et al., 2013b; Sonnleitner et al., 2013; for different results, see Kretzschmar et al., 2016; Lotz et al., 2016) and beyond a CRS task (Greiff et al., 2015b). MCS have been proposed as a tool for assessing 21st Century skills (Greiff et al., 2014) and were applied in the international large-scale study PISA to assess general problem-solving skills (OECD, 2014). They have further been proposed as training tools and evaluation instruments for these skills (e.g., Greiff et al., 2013a; Herde et al., 2016). This begs the question: how strong is the empirical evidence? Are these far-reaching conclusions and recommendations justified?

Studies provide support for the psychometric quality, especially the reliability, of the MCS approach, although scale building and some statistics have been criticized (Funke et al., 2017; Kretzschmar, 2017). Only one study so far has attempted to compare MCS and CRS. In it, Greiff et al. (2015b) argued that MCS had a higher validity than *Tailorshop* in predicting school grades. The knowledge scales assessed after exploring the system were used as predictors for the MCS. However, system-specific knowledge for *Tailorshop* after controlling the system was not assessed (Kretzschmar, 2017). Instead, control performance was used as a predictor of school grades. Control performance, however, is not a valid measure of acquired knowledge, as demonstrated in our first study. For this, additional tests are needed after controlling the system, conducted in both studies in this paper.

Minimally complex systems research also only sparingly addresses questions of construct validity related to the measures and the conclusions (i.e., generalizability; see Kretzschmar, 2015). This concerns the operationalization of CPS characteristics (i.e., the construct validity of the MCS), which was addressed in more detail above. However, limitations also exist concerning the choice of the additional instruments applied in validation studies. The construct validity of many instruments is considerably limited, causing results to be overgeneralized (cf., Shadish et al., 2002). For example, operationalizing reasoning (i.e., fluid intelligence) with a single task (e.g., the Raven matrices; Wüstenberg et al., 2012; Greiff and Fischer, 2013) is not sufficient. Construct validity is also restricted if only one task is used to measure WMC (e.g., Bühner et al., 2008; Schweizer et al., 2013). Since Spearman’s (1904), work we know that task-specific variance can be reduced only through heterogeneous operationalizations of the intended constructs. The two studies reported in this paper show how strongly the relationship between intelligence and CPS performance varies depending on the generality level of the intelligence construct (see also Kretzschmar et al., 2017). The symmetry problem was demonstrated here for the BIS, but is also evident with regard to other hierarchical intelligence models, e.g., the Three Stratum theory (Carroll, 1993, 2005), the extended Gf-Gc theory (Horn and Blankson, 2005; Horn, 2008), and the Cattell-Horn-Carroll theory (CHC theory; McGrew, 2005, 2009).

Süß and Beauducel (2011), therefore, classified every task of the most frequently used tests into the BIS, the three stratum theory, and the CHC theory to give a framework for this problem.

According to the BIS (Jäger, 1982), every intelligence task depends on at least two abilities (an operative and a content ability), i.e., every task relates to two different constructs. By extension, the interpretation in terms of only one ability is of limited validity due to unintended but reliable task-specific variance. It is either necessary to have several tasks for every construct and theory-based aggregation (Jäger, 1982, 1984) to reduce unintended variance, or the interpretation must be limited to a more specific conclusion (e.g., to numerical reasoning in our first study). The two studies presented here and many others show that these kinds of problems substantially influence the validity of conclusions in intelligence and problem solving research as well as in many other fields (Shadish et al., 2002).

In summary, the MCS approach provides solutions to psychometrics problems in CPS research, especially the reliability problem, but its validity as an indicator of CPS performance is substantially restricted. In our view, MCS are an interesting new class of problem-solving tasks, but provide few insights into complex real-world problem solving. Modifications of the MCS approach toward increased complexity (e.g., MicroFIN; Neubert et al., 2015; Kretzschmar et al., 2017) are a promising step in the right direction.

## Conclusion and Outlook

The primary aim of CPS research with CRSs (e.g., Lohhausen; Dörner et al., 1983) is ecological validity, i.e., “the validity of the empirical results as psychological statements for the real world” (Fahrenberg, 2017). In the past, many systems were “*ad hoc*” constructions by psychologists that had not been sufficiently validated, but this need not be the case. What is needed is interdisciplinary research in the form of collaboration with experts in the simulated domains. For example, Dörner collaborated with a business expert to develop *Tailorshop*. *Powerplant* was developed by Wallach (1997) together with engineers from a coal-fired power plant near Saarbrücken (Germany). *LEARN!*, a complex management simulator with more than 2000 connected variables, was originally developed by an economics research group at the University of Mannheim (Germany) as a tool for testing economic theories (Milling, 1996; Größler et al., 2000; Maier and Größler, 2000). In the version applied by Wittmann et al. (1996), participants have to manage a high-technology company competing with three others simulated by the computer. *ATC* (*Air Traffic Controller Test*; Ackerman and Kanfer, 1993) and *TRACON* (*Terminal Radar Air Control*; Ackerman, 1992) are simplified versions of vocational training simulators for professional air traffic controllers. The *Situational Awareness Real Time Assessment Tool* (*SARA-T*) was developed to measure the situational awareness of air traffic controllers working in the *NLR ATM Research Simulator* (*NARSIM*; ten Have, 1993), a system also used in expert studies (Kraemer and Süß, 2015; Kraemer, 2018). Finally, technological developments (e.g., video clips, virtual worlds; Funke, 1998) have enabled the development of

complex systems that are much more similar to real-world demands than ever before, an opportunity that should be capitalized upon in psychological research (see Dörner and Funke, 2017).

In this line of research, the ecological validity of the simulated real-world relationships is essential and must be ensured. In addition, domain-specific prior knowledge is necessary to generate hypotheses for system exploration and system control. Valid measures of the amount, type, and structure of domain-specific prior knowledge, the knowledge acquisition processes, and the acquired knowledge are necessary for understanding and measuring CPS behavior and performance. In light of all this, this line of research can help us to understand how people face the challenge of dealing with complexity and uncertainty, identify causes of failure, and detect successful strategies for reducing complexity during problem solving (e.g., Dörner, 1996; Dörner and Funke, 2017), a laborious and time-consuming but important field of research in complex decision making (cf., Gigerenzer and Gaissmaier, 2011). The research strategy of restricting complex problem solving tasks to MCS, however, leads into a cul-de-sac.

## ETHICS STATEMENT

The studies were carried out in accordance with the ethical guidelines of the German Association of Psychology with informed consent from all subjects. Considering the time when the studies were conducted and the fact that the materials and procedures were not invasive, the studies were not approved by an ethical committee.

## AUTHOR CONTRIBUTIONS

H-MS conceptualized the manuscript and conducted the first study. AK conducted the second study. H-MS and AK analyzed the data and drafted the manuscript in collaboration.

## FUNDING

The first study was supported by a grant from the Free University of Berlin's Commission for Research Promotion (FNK) to the first author and A. O. Jäger. We further acknowledge support by the Deutsche Forschungsgemeinschaft and the University of Tübingen's Open Access Publishing Fund. In addition, this research project was supported by the Postdoc Academy of the Hector Research Institute of Education Sciences and Psychology, Tübingen, funded by the Baden-Württemberg Ministry of Science, Education and the Arts.

## ACKNOWLEDGMENTS

We thank Klaus Oberauer for his helpful comments on the manuscript.



## REFERENCES

- Ackerman, P. L. (1992). Predicting individual differences in complex skill acquisition: dynamics of ability determinants. *J. Appl. Psychol.* 77, 598–614. doi: 10.1037/0021-9010.77.5.598
- Ackerman, P. L., and Kanfer, R. (1993). Integrating laboratory and field study for improving selection: development of a battery for predicting air traffic controller success. *J. Appl. Psychol.* 78, 413–432. doi: 10.1037/0021-9010.78.3.413
- Bainbridge, L. (1974). "Analysis of verbal protocols from a process control task," in *The Human Operator in Process Control*, eds E. Edwards and F. P. Lees (London: Taylor & Francis), 146–158.
- Barth, C. M., and Funke, J. (2010). Negative affective environments improve complex solving performance. *Cogn. Emot.* 24, 1259–1268. doi: 10.1080/02699930903223766
- Beckmann, J. F., and Goode, N. (2014). The benefit of being naïve and knowing it: the unfavourable impact of perceived context familiarity on learning in complex problem solving tasks. *Instr. Sci.* 42, 271–290. doi: 10.1007/s11251-013-9280-7
- Berry, D. C., and Broadbent, D. E. (1988). Interactive tasks and the implicit-explicit distinction. *Br. J. Psychol.* 79, 251–272. doi: 10.1111/j.2044-8295.1988.tb02286.x
- Berry, D. C., and Dienes, Z. (1993). *Implicit Learning. Theoretical and Empirical Issues*. Hillsdale, MI: LEA.
- Brehmer, B. (1986). "In one word: not from experience," in *Judgment and Decision Making*, eds H. R. Arkes and K. R. Hammond (Cambridge: Cambridge University Press), 705–720.
- Broadbent, D. E., Fitzgerald, P., and Broadbent, M. H. P. (1986). Implicit and explicit knowledge in the control of complex systems. *Br. J. Psychol.* 77, 33–50. doi: 10.1111/j.2044-8295.1986.tb01979.x
- Buchner, A., and Funke, J. (1993). Finite state automata: dynamic task environments in problem solving research. *Q. J. Exp. Psychol.* 46A, 83–118. doi: 10.1080/14640749308401068
- Bühner, M., Kröner, S., and Ziegler, M. (2008). Working memory, visual-spatial-intelligence and their relationship to problem-solving. *Intelligence* 36, 672–680. doi: 10.1016/j.intell.2008.03.008
- Carroll, J. B. (1993). *Human Cognitive Abilities. A Survey of Factor-Analytic Studies*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511571312
- Carroll, J. B. (2005). "The three-stratum theory of cognitive abilities," in *Contemporary Intellectual Assessment: Theories, Test, and Issues*, 2nd Edn, eds D. P. Flanagan and P. L. Harrison (New York, NY: Guilford Press), 69–76.
- Cattell, R. B. (1987). *Intelligence: Its Structure, Growth, and Action*. Amsterdam: Elsevier.
- Ceci, S. J., and Liker, J. K. (1986). A day at the races: a study of IQ, expertise, and cognitive complexity. *J. Exp. Psychol. Gen.* 115, 255–266. doi: 10.1037/0096-3445.115.3.255
- Chi, M. T. H., Glaser, R., and Farr, M. J. (1988). *The Nature of Expertise*. Hillsdale, NJ: Erlbaum.
- Conway, A. R., Kane, M. J., and Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends Cogn. Sci.* 7, 547–552. doi: 10.1016/j.tics.2003.10.005
- Csapó, B., and Molnár, G. (2017). Potential for assessing dynamic problem-solving at the beginning of higher education studies. *Front. Psychol.* 8:2022. doi: 10.3389/fpsyg.2017.02022
- Danner, D., Hagemann, D., Holt, D. V., Hager, M., Schankin, A., Wüstenberg, S., et al. (2011). Measuring performance in dynamic decision making: reliability and validity of the *Tailorshop* simulation. *J. Individ. Dif.* 32, 225–233. doi: 10.1027/1614-0001/a000055
- Deutsche Gesellschaft für Personalwesen [DGP] (1986). *Differentieller Kenntnistest (DKT). Subtest Wirtschaft [Differential Test of Knowledge: Subtest Economics]*. Hannover: DGP.
- Dörner, D. (1986). Diagnostik der operativen Intelligenz [Diagnostics of operative intelligence]. *Diagnostica* 32, 290–308.
- Dörner, D. (1996). *The Logic of Failure: Recognizing and Avoiding Error in Complex Situations*. New York, NY: Basic Books.
- Dörner, D., and Funke, J. (2017). Complex problem solving: what it is and what it is not. *Front. Psychol.* 8:1153. doi: 10.3389/fpsyg.2017.01153
- Dörner, D., and Kreuzig, H. W. (1983). Problemlösefähigkeit und Intelligenz [Problem solving ability and intelligence]. *Psychol. Rundsch.* 34, 185–192.
- Dörner, D., Kreuzig, H. W., Reither, F., and Stäudel, T. (1983). *Lohhausen. Vom Umgang mit Unbestimmtheit und Komplexität [Lohhausen. About Dealing with Uncertainty and Complexity]*. Bern: Huber.
- Dutt, V., and Gonzalez, C. (2015). Accounting for outcome and process measures in dynamic decision-making tasks through model calibration. *J. Dyn. Decis. Mak.* 1, 1–10. doi: 10.11588/jddm.2015.1.17663
- Engelhart, M., Funke, J., and Sager, S. (2017). A web-based feedback study on optimization-based training and analysis of human decision making. *J. Dyn. Decis. Mak.* 3, 1–23. doi: 10.11588/jddm.2017.1.34608
- Fahrenberg, J. (2017). "Ökologische Validität [ecological validity]," in *Dorsch - Lexikon der Psychologie*, ed. H. Wirz (Bern: Huber), 1202.
- Fishbein, M., and Ajzen, I. (1974). Attitudes towards objects as predictors of single and multiple behavioral criteria. *Psychol. Rev.* 81, 59–74. doi: 10.1037/h0035872
- Frensch, P. A., and Funke, J. (1995). "Definitions, traditions, and a general framework for understanding complex problem solving," in *Complex Problem Solving. The European Perspective*, eds P. A. Frensch and J. Funke (Hillsdale, NJ: Lawrence Erlbaum Associates), 3–25.
- Funke, J. (1983). Einige Bemerkungen zu Problemen der Problemlöseforschung oder: Ist Testintelligenz doch ein Prädiktor? [Issues in problem solving research: is test intelligence a predictor after all?]. *Diagnostica* 29, 283–302.
- Funke, J. (1985). Steuerung dynamischer Systeme durch Aufbau und Anwendung subjektiver Kausalmodelle [Control of dynamic systems by building up and using subjective causal models]. *Z. Psychol.* 193, 435–457.
- Funke, J. (1992). *Wissen über dynamische Systeme: Erwerb, Repräsentation und Anwendung [Knowledge About Dynamic Systems: Acquisition, Representation, and use]*. Berlin: Springer. doi: 10.1007/978-3-642-77346-4
- Funke, J. (1993). "Microworlds based on linear equation systems: a new approach to complex problem solving and experimental results," in *The Cognitive Psychology of Knowledge*, eds G. Strube and K.-F. Wender (Amsterdam: Elsevier), 313–330.
- Funke, J. (1998). Computer-based testing and training with scenarios from complex problem solving research: advantages and disadvantages. *Int. J. Sel. Assess.* 6, 90–96. doi: 10.1111/1468-2389.00077
- Funke, J. (2006). "Komplexes Problemlösen," in *Denken und Problemlösen (Enzyklopädie der Psychologie, Serie II Kognition, Bd. 8)*, ed. J. Funke (Göttingen: Hogrefe), 375–445.
- Funke, J. (2014). Analysis of minimal complex systems and complex problem solving require different forms of causal cognition. *Front. Psychol.* 5:739. doi: 10.3389/fpsyg.2014.00739
- Funke, J., Fischer, A., and Holt, V. D. (2017). When less is less: solving multiple simple problems is not complex problem solving—a comment on Greiff et al. (2015). *J. Intell.* 5:5. doi: 10.3390/jintelligence5010005
- Funke, U. (1995). "Using complex problem solving tasks in personnel selection and training," in *Complex Problem Solving. The European Perspective*, eds P. A. Frensch and J. Funke (Hillsdale NJ: Erlbaum), 219–240.
- Gigerenzer, G., and Gaissmaier, W. (2011). Heuristic decision making. *Annu. Rev. Psychol.* 62, 451–482. doi: 10.1146/annurev-psych-120709-145346
- Gonzalez, C., and Dutt, V. (2011). A generic dynamic control task for behavioral research and education. *Comput. Hum. Behav.* 27, 1904–1914. doi: 10.1016/j.chb.2011.04.015
- Gonzalez, C., Lerch, J. F., and Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cogn. Sci.* 27, 591–635. doi: 10.1016/S0364-0213(03)00031-4
- Goode, N., and Beckmann, J. F. (2010). You need to know: there is a causal relationship between structural knowledge and control performance in complex problem solving tasks. *Intelligence* 38, 345–352. doi: 10.1016/j.intell.2010.01.001
- Goode, N., and Beckmann, J. F. (2016). With a little help . . . on the role of guidance in the acquisition and utilisation of knowledge in the control of complex, dynamic systems. *J. Dyn. Decis. Mak.* 2:4. doi: 10.11588/jddm.2016.1.33346
- Greiff, S., and Fischer, A. (2013). Der Nutzen einer komplexen Problemlösekompetenz: Theoretische Überlegungen und empirische Befunde [The value of complex problem solving competency: theoretical considerations and empirical results]. *Z. Pädagog. Psychol.* 27, 27–39. doi: 10.1024/1010-0652/a000086
- Greiff, S., Fischer, A., Stadler, M., and Wüstenberg, S. (2015a). Assessing complex problem-solving skills with multiple complex systems. *Think. Reason.* 21, 356–382. doi: 10.1080/13546783.2014.989263

- Greiff, S., Kretzschmar, A., Müller, J. C., Spinath, B., and Martin, R. (2014). The computer-based assessment of complex problem solving and how it is influenced by students' information and communication technology literacy. *J. Educ. Psychol.* 106, 666–680. doi: 10.1037/a0035426
- Greiff, S., Krkovic, K., and Hautamäki, J. (2016). The prediction of problem-solving assessed via microworlds a study on the relative relevance of fluid reasoning and working memory. *Eur. J. Psychol. Assess.* 32, 298–306. doi: 10.1027/1015-5759/a000263
- Greiff, S., Stadler, M., Sonnleitner, P., Wolff, C., and Martin, R. (2015b). Sometimes less is more: comparing the validity of complex problem solving measures. *Intelligence* 50, 100–113. doi: 10.1016/j.intell.2015.02.007
- Greiff, S., Wüstenberg, S., and Funke, J. (2012). Dynamic problem solving: a new assessment perspective. *Appl. Psychol. Meas.* 36, 189–213. doi: 10.1177/0146621612439620
- Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F., and Funke, J. (2013a). Computer-based assessment of Complex Problem Solving: concept, implementation, and application. *Educ. Technol. Res. Dev.* 61, 407–421. doi: 10.1007/s11423-013-9301-x
- Greiff, S., Wüstenberg, S., Molnar, G., Fischer, A., Funke, J., and Csapo, B. (2013b). Complex problem solving in educational contexts—something beyond g: concept, assessment, measurement invariance, and construct validity. *J. Educ. Psychol.* 105, 364–379. doi: 10.1037/a0031856
- Größler, A., Maier, F. H., and Milling, P. M. (2000). Enhancing learning capabilities by providing transparency in business simulators. *Simul. Gaming* 31, 257–278. doi: 10.1177/104687810003100209
- Gustafsson, J.-E., and Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behav. Res.* 28, 407–434. doi: 10.1207/s15327906mbr2804\_2
- Haynes, S. N., Richard, D. C. S., and Kubany, E. S. (1995). Content validity in psychological assessment: a functional approach to concepts and methods. *Psychol. Assess.* 7, 238–247. doi: 10.1037/1040-3590.7.3.238
- Herde, C. N., Wüstenberg, S., and Greiff, S. (2016). Assessment of complex problem solving: what we know and what we don't know. *Appl. Meas. Educ.* 29, 265–277. doi: 10.1080/08957347.2016.1209208
- Hesse, F. W. (1982). Effekte des semantischen Kontextes auf die Bearbeitung komplexer Probleme [Effect of semantic context on the solution of complex problems]. *Z. Exp. Angew. Psychol.* 29, 62–91.
- Horn, J. L. (2008). "Spearman, g, expertise, and the nature of human cognitive capability," in *Extending Intelligence: Enhancement and New Constructs*, eds P. C. Kyllonen, R. D. Roberts, and L. Stankov (New York, NY: Lawrence Erlbaum Associates), 185–230.
- Horn, J. L., and Blankson, N. (2005). "Foundations for better understanding of cognitive abilities," in *Contemporary Intellectual Assessment: Theories, Tests, and Issues*, 2nd Edn, eds D. P. Flanagan and P. I. Harrison (New York, NY: Guilford Press), 41–68.
- Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen. Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells [Multimodal classification of intellectual performance. Experimental development of a descriptive intelligence structure model]. *Diagnostica* 28, 195–226.
- Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven [Intelligence structure research: competing models, new developments, perspectives]. *Psychol. Rundsch.* 35, 21–35.
- Jäger, A. O., Süß, H.-M., and Beauducel, A. (1997). *Test für das Berliner Intelligenzstrukturmodell. BIS-Test. Form 4 [Test for the Berlin Intelligence Structure Model]*. Göttingen: Hogrefe.
- Jensen, A. R., and Wang, L.-J. (1994). What is a good g? *Intelligence* 18, 231–258. doi: 10.1016/0160-2896(94)90029-9
- Kerlinger, F. N., and Pedhazur, E. J. (1973). *Multiple Regression in Behavioral Research*. New York, NY: Holt, Rinehart and Winston.
- Kersting, M. (2001). Zur Konstrukt- und Kriteriumsvalidität von Problemlösenszenarien anhand der Vorhersage von Vorgesetztenurteilen über die berufliche Bewährung [On the construct and criterion validity of problem-solving scenarios based on the prediction of supervisor assessment of job performance]. *Diagnostica* 47, 67–76. doi: 10.1026//0012-1924.47.2.67
- Kersting, M., and Süß, H.-M. (1995). Kontentvalide Wissensdiagnostik und Problemlösen: Zur Entwicklung, testtheoretischen Begründung und empirischen Bewährung eines problemspezifischen Diagnoseverfahrens [Content-valid diagnosis of knowledge and problem-solving: development, test theory justification, and empirical validation of a new problem-specific test]. *Z. Pädagog. Psychol.* 9, 83–93.
- King, J., and Just, M. A. (1991). Individual differences in syntactic processing: the role of working memory. *J. Mem. Lang.* 30, 580–602. doi: 10.1016/0749-596X(91)90027-H
- Klauer, K. J. (1984). Kontentvalidität. [Content validity]. *Diagnostica* 30, 1–23.
- Kluge, A. (2008b). What you train is what you get? Task requirements and training methods in complex problem-solving. *Comput. Hum. Behav.* 24, 284–308. doi: 10.1016/j.chb.2007.01.013
- Kluge, A. (2008a). Performance assessments with microworlds and their difficulty. *Appl. Psychol. Meas.* 32, 156–180. doi: 10.1177/0146621607300015
- Kluwe, R. H., Misiak, C., and Haider, H. (1991). "The control of complex systems and performance in intelligence tests," in *Intelligence: Reconceptualization and Measurement*, ed. H. Rowe (Hillsdale: Lawrence Erlbaum Associates).
- Kotter, J. P. (1982). What effective general managers really do. *Harv. Bus. Rev.* 60, 156–167.
- Kraemer, J. (2018). *Die Lücke im Entscheidungsprozess. Die Bedeutsamkeit von Situationsbewusstsein und Optionsgenerierung für die Leistung von Fluglotsen [The Gap in Decision Making. The Significance of Situation Awareness and Option Generation for Air Traffic Controller Performance]*. Köln: Deutsches Zentrum für Luft- und Raumfahrt e. V.
- Kraemer, J., and Süß, H.-M. (2015). Real time validation of online situation awareness questionnaires in simulated approach air traffic control. *Procedia Manuf.* 3, 3152–3159. doi: 10.1016/j.promfg.2015.07.864
- Kretzschmar, A. (2015). *Konstruktvalidität des komplexen Problemlösens Unter Besonderer Berücksichtigung Moderner Diagnostischer Ansätze [Construct Validity of Complex Problem Solving With Particular Focus on Modern Assessment Approaches]*. Doctoral dissertation, University of Luxembourg, Luxembourg.
- Kretzschmar, A. (2017). Sometimes less is not enough: a commentary on Greiff et al. (2015). *J. Intell.* 5:4. doi: 10.3390/jintelligence5010004
- Kretzschmar, A., Hacıtrjana, L., and Rasevska, M. (2017). Re-evaluating the psychometric properties of MicroFIN: a multidimensional measurement of complex problem solving or a unidimensional reasoning test? *Psychol. Test Assess. Model.* 59, 157–182.
- Kretzschmar, A., Neubert, J. C., and Greiff, S. (2014). Komplexes Problemlösen, schulfachliche Kompetenzen und ihre Relation zu Schulnoten [Complex problem solving, school competencies and their relation to school grades]. *Z. Pädagog. Psychol.* 28, 205–215. doi: 10.1024/1010-0652/a000137
- Kretzschmar, A., Neubert, J. C., Wüstenberg, S., and Greiff, S. (2016). Construct validity of complex problem solving: a comprehensive view on different facets of intelligence and school grades. *Intelligence* 54, 55–69. doi: 10.1016/j.intell.2015.11.004
- Kretzschmar, A., and Süß, H.-M. (2015). A study on the training of complex problem solving competence. *J. Dyn. Decis. Mak.* 1, 1–14. doi: 10.11588/jddm.2015.1.15455
- Kröner, S. (2001). *Intelligenzdiagnostik per Computersimulation [Intelligence assessment via computer simulation]*. Münster: Waxmann.
- Kröner, S., Plass, J. L., and Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence* 33, 347–368. doi: 10.1016/j.intell.2005.03.002
- Krumm, V., and Seidel, G. (1970). *Wirtschaftslehretest [Economics Test]*. Weinheim: Beltz.
- Kyllonen, P. C., and Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence* 14, 389–433. doi: 10.1016/S0160-2896(05)80012-1
- Leutner, D. (2002). The fuzzy relationship of intelligence and problem solving in computer simulations. *Comput. Hum. Behav.* 18, 685–697. doi: 10.1016/s0747-5632(02)00024-9
- Lotz, C., Scherer, R., Greiff, S., and Sparfeldt, J. R. (2017). Intelligence in action – Effective strategic behaviors while solving complex problems. *Intelligence* 64, 98–112. doi: 10.1016/j.intell.2017.08.002
- Lotz, C., Sparfeldt, J. R., and Greiff, S. (2016). Complex problem solving in educational contexts – Still something beyond a "good g"? *Intelligence* 59, 127–138. doi: 10.1016/j.intell.2016.09.001

- Maier, F. H., and Größler, A. (2000). What are we talking about? - A taxonomy of computer simulations to support learning. *Syst. Dyn. Rev.* 16, 135–148. doi: 10.1002/1099-1727(200022)16:2<135::AID-SDR193>3.0.CO;2-P
- McGrew, K. S. (2005). “The Cattell-Horn-Carroll theory of cognitive abilities,” in *Contemporary Intellectual Assessment: Theories, Test, and Issues*, 2nd Edn, eds D. P. Flanagan and P. L. Harrison (New York, NY: Guilford Press), 136–181.
- McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: standing on the shoulders of the giants of psychometric intelligence research. *Intelligence* 37, 1–10. doi: 10.1016/j.intell.2008.08.004
- Milling, P. M. (1996). Modeling innovation processes for decision support and management simulation. *Syst. Dyn. Rev.* 12, 211–234. doi: 10.1002/(SICI)1099-1727(199623)12:3<211::AID-SDR105>3.0.CO;2-8
- Mintzberg, H. (1973). *The Nature of Managerial Work*. New York, NY: Harper & Row.
- Morris, N. M., and Rouse, W. B. (1985). The effects of type of knowledge upon human problem solving in a process control task. *IEEE Trans. Syst. Man Cybern.* 15, 698–707. doi: 10.1109/TSMC.1985.6313453
- Neubert, J. C., Kretzschmar, A., Wüstenberg, S., and Greiff, S. (2015). Extending the assessment of complex problem solving to finite state automata: embracing heterogeneity. *Eur. J. Psychol. Assess.* 31, 181–194. doi: 10.1027/1015-5759/a000224
- Oberauer, K., Süß, H.-M., Wilhelm, O., and Wittmann, W. W. (2003). The multiple faces of working memory: storage, processing, supervision, and coordination. *Intelligence* 31, 167–193. doi: 10.1016/S0160-2896(02)00115-0
- Oberauer, K., Süß, H.-M., Wilhelm, O., and Wittmann, W. W. (2008). Which working memory functions predict intelligence? *Intelligence* 36, 641–652. doi: 10.1016/j.intell.2008.01.007
- OECD (2014). *Pisa 2012 Results: Creative Problem Solving: Students' Skills in Tackling Real-Life Problems (Volume V)*. Paris: OECD Publishing. doi: 10.1787/9789264208070-en
- Putz-Osterloh, W. (1981). Über die Beziehung zwischen Testintelligenz und Problemlöseerfolg [On the relationship between test intelligence and problem solving success]. *Z. Psychol.* 189, 79–100.
- Rigas, G., Carling, E., and Brehmer, B. (2002). Reliability and validity of performance measures in microworlds. *Intelligence* 30, 463–480. doi: 10.1016/S0160-2896(02)00121-6
- Ryan, K. J. (2006). *The Relationship Between Complex Problem Solving and Intelligence: An Analysis of Three Computer Simulated Scenarios*. Doctoral dissertation, University of Sydney, Sydney.
- Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.
- Scherer, R. (2015). Is it time for a new measurement approach? A closer look at the assessment of cognitive adaptability in complex problem solving. *Front. Psychol.* 6:1664. doi: 10.3389/fpsyg.2015.01664
- Scherer, R., and Tiemann, R. (2014). Measuring students' progressions in scientific problem solving: a psychometric approach. *Procedia Soc. Behav. Sci.* 112, 87–96. doi: 10.1016/j.sbspro.2014.01.1142
- Schmid, J., and Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika* 22, 53–61. doi: 10.1007/BF02289209
- Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *Am. Psychol.* 47, 1173–1181. doi: 10.1037/0003-066X.47.10.1173
- Schmidt, F. L., Hunter, J. E., and Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *J. Appl. Psychol.* 71, 432–439. doi: 10.1037/0021-9010.71.3.432
- Schoppek, W., and Fischer, A. (2015). Complex problem solving-single ability or complex phenomenon? *Front. Psychol.* 6:1669. doi: 10.3389/fpsyg.2015.01669
- Schweizer, F., Wüstenberg, S., and Greiff, S. (2013). Validity of the MicroDYN approach: complex problem solving predicts school grades beyond working memory capacity. *Learn. Individ. Dif.* 24, 42–52. doi: 10.1016/j.lindif.2012.12.011
- Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton-Mifflin.
- Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., et al. (2012). The genetics lab: acceptance and psychometric characteristics of a computer-based microworld assessing complex problem solving. *Psychol. Test Assess. Model.* 54, 54–72. doi: 10.1037/e578442014-045
- Sonnleitner, P., Keller, U., Martin, R., and Brunner, M. (2013). Students' complex problem-solving abilities: their structure and relations to reasoning ability and educational success. *Intelligence* 41, 289–305. doi: 10.1016/j.intell.2013.05.002
- Spearman, C. (1904). “General intelligence”, objectively determined and measured. *Am. J. Psychol.* 15, 201–293. doi: 10.2307/1412107
- Spering, M., Wagener, D., and Funke, J. (2005). The role of emotions in complex problem-solving. *Cogn. Emot.* 19, 1252–1261. doi: 10.1080/02699930500304886
- Stadler, M. J., Becker, N., Gödker, M., Leutner, D., and Greiff, S. (2015). Complex problem solving and intelligence: a meta-analysis. *Intelligence* 53, 92–101. doi: 10.1016/j.intell.2015.09.005
- Stadler, M. J., Becker, N., Greiff, S., and Spinath, F. M. (2016). The complex route to success: complex problem-solving skills in the prediction of university success. *High. Educ. Res. Dev.* 35, 365–379. doi: 10.1080/07294360.2015.1087387
- Stankov, L. (2017). Overemphasized “g”. *J. Intell.* 5:33. doi: 10.3390/jintelligence5040033
- Süß, H.-M. (1996). *Intelligenz, Wissen und Problemlösen. Kognitive Voraussetzungen für erfolgreiches Handeln bei computersimulierten Problemen [Intelligence, Knowledge, and Problem Solving: Cognitive Prerequisites of Successful Performance in Computer-Simulated Problems]*. Lehr- und Forschungstexte Psychologie. Göttingen: Hogrefe.
- Süß, H.-M. (1999). Intelligenz und komplexes Problemlösen: Perspektiven für eine Kooperation zwischen differentiell-psychometrischer und kognitionspsychologischer Forschung [Intelligence and complex problem solving: perspectives on the cooperation between differential-psychometric and cognitive research methods]. *Psychol. Rundsch.* 50, 220–228. doi: 10.1026//0033-3042.50.4.220
- Süß, H.-M. (2001). “The predictive validity of reasoning and g in complex problem solving,” in *Paper Presented at the ISSID 2001 Conference*, Edinburgh.
- Süß, H.-M., and Beauducel, A. (2005). “Faceted models of intelligence,” in *Understanding and Measuring Intelligence*, eds O. Wilhelm and R. Engle (Thousand Oaks, CA: Sage), 313–332.
- Süß, H.-M., and Beauducel, A. (2011). “Intelligenztests und ihre Bezüge zu Intelligenztheorien. [Intelligence tests and their relationships to theories of intelligence],” in *Leistungs-, Intelligenz- und Verhaltensdiagnostik (Enzyklopädie der Psychologie, Serie Psychologische Diagnostik, Bd. 3)*, eds L. F. Hornke, M. Amelang, and M. Kersting (Göttingen: Hogrefe), 97–234.
- Süß, H.-M., and Beauducel, A. (2015). Modeling the construct validity of the Berlin intelligence structure model. *Estud. Psicol.* 32, 13–25. doi: 10.1590/0103-166X2015000100002
- Süß, H.-M., and Faulhaber, J. (1990). *Berliner Version der Schneiderwerkstatt. PC-Simulationsprogramm [Berlin Version of the Tailorshop]*. Berlin: Freie Universität Berlin, Fachbereich Erziehungs- und Unterrichtswissenschaften, Institut für Psychologie.
- Süß, H.-M., Kersting, M., and Oberauer, K. (1991). Intelligenz und Wissen als Prädiktoren für Leistungen bei computersimulierten komplexen Problemen [Intelligence and knowledge as predictors of performance in solving complex computer-simulated problems]. *Diagnostica* 37, 334–352.
- Süß, H.-M., Kersting, M., and Oberauer, K. (1993a). Zur Vorhersage von Steuerungsleistungen an computersimulierten Systemen durch Wissen und Intelligenz [On the predictability of control performance on computer-simulated systems by knowledge and intelligence]. *Z. Differ. Diagnostische Psychol.* 14, 189–203.
- Süß, H.-M., Oberauer, K., and Kersting, M. (1993b). Intellektuelle Fähigkeiten und die Steuerung komplexer Systeme [Intelligence and control performance on computer-simulated systems]. *Spr. Kognition* 12, 83–97.
- Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O., and Schulze, R. (2002). Working-memory capacity explains reasoning ability - And a little bit more. *Intelligence* 30, 261–288. doi: 10.1016/S0160-2896(01)00100-3
- Swanson, L., and Kim, K. (2007). Working memory, short-term memory, and naming speed as predictors of children's mathematical performance. *Intelligence* 35, 151–168. doi: 10.1016/j.intell.2006.07.001
- ten Have, J. M. (1993). The development of the NLR ATC Research Simulator (Narsim): design philosophy and potential for ATM research. *Simul. Pract. Theory* 1, 31–39. doi: 10.1016/0928-4869(93)90009-F
- Vollmeyer, R., Burns, B. D., and Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cogn. Sci.* 20, 75–100. doi: 10.1207/s15516709cog2001\_3

- Wagener, D. (2001). *Psychologische Diagnostik mit komplexen Szenarios - Taxonomie, Entwicklung, Evaluation [Psychological Assessment with Complex Scenarios - Taxonomy, Development, Evaluation]*. Lengerich: Pabst Science Publishers.
- Wagener, D., and Wittmann, W. W. (2002). Personalarbeit mit dem komplexen Szenario FSYS [Human resource management using the complex scenario FSYS]. *Z. Personalpsychologie* 1, 80–93. doi: 10.1026//1617-6391.1.2.80
- Wallach, D. (1997). *Kognitionswissenschaftliche Analysen komplexer Problemlöseprozesse [Cognitive Science Analyses of Complex Problem Solving Processes]*. Wiesbaden: Westdeutscher Verlag.
- Wittmann, W. W. (1988). "Multivariate reliability theory. Principles of symmetry and successful validation strategies," in *Handbook of Multivariate Experimental Psychology*, eds R. B. Cattell and J. R. Nesselroade (New York, NY: Plenum), 505–560. doi: 10.1007/978-1-4613-0893-5\_16
- Wittmann, W. W., and Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Syst. Res. Behav. Sci.* 21, 393–409. doi: 10.1002/sres.653
- Wittmann, W. W., and Süß, H.-M. (1999). "Investigating the paths between working memory, intelligence, knowledge, and complex problem-solving performances via Brunswik symmetry," in *Learning and Individual Differences: Process, Trait and Content Determinants*, eds P. L. Ackerman, P. C. Kyllonen, and R. D. Roberts (Washington, DC: APA), 77–104.
- Wittmann, W. W., Süß, H.-M., and Oberauer, K. (1996). *Determinanten komplexer Problemlösens [Determinants of Complex Problem Solving]*. Research Report No. 9. Mannheim: Universität Mannheim.
- Wolfe, J., and Roberts, C. R. (1986). The external validity of a business management game: a five-year longitudinal study. *Simul. Games* 17, 45–59. doi: 10.1177/0037550086171004
- Wüstenberg, S., Greiff, S., and Funke, J. (2012). Complex problem solving - More than reasoning? *Intelligence* 40, 1–14. doi: 10.1016/j.intell.2011.11.003
- Zech, A., Bühner, M., Kröner, S., Heene, M., and Hilbert, S. (2017). The impact of symmetry: explaining contradictory results concerning working memory, reasoning, and complex problem solving. *J. Intell.* 5:22. doi: 10.3390/jintelligence5020022

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Süß and Kretzschmar. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# The Efficacy and Development of Students' Problem-Solving Strategies During Compulsory Schooling: Logfile Analyses

Gyöngyvér Molnár<sup>1\*</sup> and Benő Csapó<sup>2</sup>

<sup>1</sup> Department of Learning and Instruction, University of Szeged, Szeged, Hungary, <sup>2</sup> MTA-SZTE Research Group on the Development of Competencies, University of Szeged, Szeged, Hungary

## OPEN ACCESS

### Edited by:

Wolfgang Schoppek,  
University of Bayreuth, Germany

### Reviewed by:

J. F. Beckmann,  
Durham University, United Kingdom  
Ronny Scherer,  
Centre for Educational Measurement,  
Faculty of Educational Sciences,  
University of Oslo, Norway

### \*Correspondence:

Gyöngyvér Molnár  
gymolnar@edpsy.u-szeged.hu

### Specialty section:

This article was submitted to  
Cognitive Science,  
a section of the journal  
Frontiers in Psychology

**Received:** 26 April 2017

**Accepted:** 23 February 2018

**Published:** 09 March 2018

### Citation:

Molnár G and Csapó B (2018) The  
Efficacy and Development of  
Students' Problem-Solving Strategies  
During Compulsory Schooling: Logfile  
Analyses. *Front. Psychol.* 9:302.  
doi: 10.3389/fpsyg.2018.00302

The purpose of this study was to examine the role of exploration strategies students used in the first phase of problem solving. The sample for the study was drawn from 3<sup>rd</sup>- to 12<sup>th</sup>-grade students (aged 9–18) in Hungarian schools ( $n = 4,371$ ). Problems designed in the MicroDYN approach with different levels of complexity were administered to the students via the eDia online platform. Logfile analyses were performed to ascertain the impact of strategy use on the efficacy of problem solving. Students' exploration behavior was coded and clustered through Latent Class Analyses. Several theoretically effective strategies were identified, including the vary-one-thing-at-a-time (VOTAT) strategy and its sub-strategies. The results of the analyses indicate that the use of a theoretically effective strategy, which extract all information required to solve the problem, did not always lead to high performance. Conscious VOTAT strategy users proved to be the best problem solvers followed by non-conscious VOTAT strategy users and non-VOTAT strategy users. In the primary school sub-sample, six qualitatively different strategy class profiles were distinguished. The results shed new light on and provide a new interpretation of previous analyses of the processes involved in complex problem solving. They also highlight the importance of explicit enhancement of problem-solving skills and problem-solving strategies as a tool for knowledge acquisition in new contexts during and beyond school lessons.

**Keywords:** complex problem solving, logfile analyses, exploration strategies, VOTAT strategies, latent class profiles

## INTRODUCTION

Computer-based assessment has presented new challenges and opportunities in educational research. A large number of studies have highlighted the importance and advantages of technology-based assessment over traditional paper-based testing (Csapó et al., 2012). Three main factors support and motivate the use of technology in educational assessment: (1) the improved efficiency and greater measurement precision in the already established assessment domains (e.g., Csapó et al., 2014); (2) the possibility of measuring constructs that would be impossible to

measure by other means (e.g., Complex Problem Solving (CPS)<sup>1</sup>; see Greiff et al., 2012, 2013); and (3) the opportunity of logging and analyzing not only observed variables, but metadata as well (Lotz et al., 2017; Tóth et al., 2017; Zoanetti and Griffin, 2017). Analyzing logfiles may contribute to a deeper and better understanding of the phenomenon under examination. Logfile analyses can provide answers to research questions which could not be answered with traditional assessment techniques.

This study focuses on problem solving, especially on complex problem solving (CPS), which reflects higher-order cognitive processes. Previous research identified three different ways to measure CPS competencies: (1) Microworlds (e.g., Gardner and Berry, 1995), (2) formal frameworks (Funke, 2001, 2010) and (3) minimal complex systems (Funke, 2014). In this paper, the focus is on the MicroDYN approach, which is a specific form of complex problem solving (CPS) in interactive situations using minimal complex systems (Funke, 2014). Recent analyses provide both a new theory and data-based evidence for a global understanding of different problem-solving strategies students employ or could employ in a complex problem-solving environment based on minimal complex systems.

The problem scenarios within the MicroDYN approach consist of a small number of variables and causal relations. From the perspective of the problem solver, solving a MicroDYN problem requires a sequence of continuous activities, in which the outcome of one activity is the input for the next. First, students interact with the simulated system, set values for the input variables, and observe the impacts of these settings on the target (dependent) variable. Then, they plot their conclusion about the causal relationships between the input and output variables on a graph (Phase 1). Next, they manipulate the independent variables again to set their values so that they result in the required values for the target variables (Phase 2).

When it comes to gathering information about a complex problem, as in the MicroDYN scenarios, there may be differences between the exploration strategies in terms of efficacy. Some of them may be more useful for generating knowledge about the system. Tschirgi (1980) identified different exploration strategies. When control of variables strategies (Greiff et al., 2014) were explored, findings showed that the vary-one-thing-at-a-time (VOTAT, Tschirgi, 1980; Funke, 2014) was the most effective strategy for identifying causal relations between the input and output variables in a minimal complex system (Fischer et al., 2012). Participants who employed this strategy tended to acquire more structural knowledge than those who used other strategies (Vollmeyer et al., 1996; Kröner et al., 2005). With the VOTAT strategy, the problem solver systematically varies only one input variable, while the others remain unchanged. This way, the effect of the variable that has just been changed can be observed directly by monitoring the changes in the output variables. There exist several types of VOTAT strategies.

Using this approach—defining the effectiveness of a strategy on a conceptual level, independently of empirical effectiveness—we developed a labeling system and a mathematical model based on all theoretically effective strategies. Thus, effectiveness was defined and linked to the amount of information extracted. An exploration strategy was defined as theoretically effective if the problem solver was able to extract all the information needed to solve the problem, independently of the application level of the information extracted and of the final achievement. We split the effectiveness of the exploration strategy and the usage and application of the information extracted to be able to solve the problem and control the system with respect to the target values based on the causal knowledge acquired. Systematicity was defined on the level of effectiveness based on the amount of information extracted and on the level of awareness based on the implementation of systematicity in time.

Students' actions were logged and coded according to our input behavior model and then clustered for comparison. We were able to distinguish three different VOTAT strategies and two successful non-VOTAT ones. We empirically tested awareness of the input behavior used in time. Awareness of strategy usage was analyzed by the sequence of the trials used, that is, by the systematicity of the trials used in time. We investigated the effectiveness of and differences in problem-solving behavior between three age groups by conducting latent class analyses to explore and define patterns in qualitatively different VOTAT strategy uses.

Although the assessment of problem solving within the MicroDYN approach is a relatively new area of research, its processes have already been studied in a number of different contexts, including a variety of educational settings with several age groups. Our cross-sectional design allows us to describe differences between age groups and outline the developmental tendencies of input behavior and strategy use among children in the age range covered by our data collection.

## REASONING STRATEGIES IN COMPLEX PROBLEM SOLVING

Problem-solving skills have been among the most extensively studied transversal skills over the last decade; they have been investigated in the most prominent comprehensive international large-scale assessments today (e.g., OECD, 2014). The common aspects in the different theoretical models are that a problem is characterized by a gap between the current state and the goal state with no immediate solution available (Mayer and Wittrock, 1996).

Parallel to the definition of the so-called twenty first-century skills (Griffin et al., 2012), recent research on problem solving disregards content knowledge and domain-specific processes. The reason for this is that understanding the structure of unfamiliar problems is more effective when it relies on abstract representation schemas and metacognitive strategies than on specifically relevant example problems (Klahr et al., 2007). That is, the focus is more on assessing domain-general

<sup>1</sup>With regard to terminology, please note that different terms are used for the subject at hand (e.g., complex problem solving, dynamic problem solving, interactive problem solving and creative problem solving). In this paper, we use the modifier “complex” (see Csapó and Funke, 2017; Dörner and Funke, 2017).

problem-solving strategies (Molnár et al., 2017), such as complex problem solving, which can be used to solve novel problems, even those arising in interactive situations (Molnár et al., 2013).

Logfile analyses make it possible to divide the continuum of a problem-solving process into several scoreable phases by extracting information from the logfile that documents students' problem-solving behavior. In our case, latent class analysis extracts information from the file that logs students' interaction with the simulated system at the beginning of the problem-solving process. The way students manipulate the input (independent) variables represents their reasoning strategy. Log data, on the one hand, make it possible to analyze qualitative differences in these strategies and then their efficiency in terms of how they generate knowledge resulting in the correct plotting of the causal relationship in Phase 1 and then the proper setting to reach the required target value in Phase 2. On the other hand, qualitative strategy data can be quantified, and an alternative scoring system can be devised.

From the perspective of the traditional psychometric approach and method of scoring, these problems form a test task consisting of two scoreable items. The first phase is a knowledge acquisition process, where scores are assigned based on how accurately the causal relationship was plotted. The second phase is knowledge application, where the correctness of the value for the target variable is scored. Such scoring based on two phases of solving MicroDYN problems has been used in a number of previous studies (e.g., Greiff et al., 2013, 2015; Wüstenberg et al., 2014; Csapó and Molnár, 2017; Greiff and Funke, 2017).

To sum up, there is great potential to investigate and cluster the problem-solving behavior and exploration strategy usage of the participants at the beginning of the problem-solving process and correlate the use of a successful exploration strategy with the model-building solution (achievement in Phase 1) observed directly in these simulated problem scenarios. Using logfile analyses (Greiff et al., 2015), the current article wishes to contribute insights into students' approaches to explore and solve problems related to minimal complex systems. By addressing research questions on the problem-solving strategies used, the study aims to understand students' exploration behavior in a complex problem-solving environment and the underlying causal relations. In this study, we show that such scoring can be developed through latent class analysis and that this alternative method of scoring may produce more reliable tests. Furthermore, such scoring can be automated and then employed in a large-scale assessment.

There are two major theoretical approaches to cognition relevant to our study; both offer general principles to interpret cognitive development beyond the narrower domain of problem solving. Piaget proposed the first comprehensive theory to explain the development of children's thinking as a sequence of four qualitatively different stages, the formal operational stage being the last one (Inhelder and Piaget, 1958), while the information processing approach describes human cognition by using terms and analogies borrowed from computer science. The information processing paradigm was not developed into an original developmental theory; it was rather aimed at reinterpreting and extending Piaget's theory (creating several

Neo-Piagetian models) and synthesizing the main ideas of the two theoretical frameworks (Demetriou et al., 1993; Siegler, 1999). One of the focal points of these models is to explain the development of children's scientific reasoning, or, more closely, the way children understand how scientific experiments can be designed and how causal relationships can be explored by systematically changing the values of (independent) variables and observing their impact on other (target) variables.

From the perspective of the present study, the essential common element of cognitive developmental research is the control of variables strategy. Klahr and Dunbar (1988) distinguished two related skills in scientific thinking, hypothesis formation and experimental design, and they integrated these skills into a coherent model for a process of scientific discovery. The underlying assumption is that knowledge acquisition requires an iterative process involving both. System control as knowledge application tends to include both processes, especially when acquired knowledge turns out to be insufficient or dysfunctional (J. F. Beckmann, personal communication, August 16, 2017). Furthermore, they separated the processes of rule induction and problem solving, defining the latter as a search in a space of rules (Klahr and Dunbar, 1988, p. 5).

de Jong and van Joolingen (1998) provided an overview of studies in scientific discovery learning with computer simulations. They concluded that a number of specific skills are needed for successful discovery, like systematic variation of variable values, which is in a focus of the present paper, and the use of high-quality heuristics for experimentation. They identified several characteristic problems in the discovery process and stressed that learners often have trouble interpreting data.

In one of the earliest systematic studies of students' problem-solving strategies, Vollmeyer et al. (1996) explored the impact of strategy systematicity and effectiveness on complex problem-solving performance. Based on previous studies, they distinguished the VOTAT strategy from other possible strategies [Change All (CA) and Heterogeneous (HT) other strategies], as VOTAT allows systematic exploration of the behavior of a system and a disconfirmation of hypotheses. In one of their experiments, they examined the hypothesis that VOTAT was more effective for acquiring knowledge than less systematic strategies. According to the results, the 36 undergraduate students had clearly shown strategy development. After interacting with the simulated system in several rounds, they tended to use the VOTAT strategy more frequently. In a second experiment, it was also demonstrated that goal specificity influences strategy use as well (Vollmeyer et al., 1996).

Beckmann and Goode (2014) analyzed the systematicity in exploration behavior in a study involving 80 first-year psychology students and focusing on the semantic context of a problem and its effect on the problem solvers' behavior in complex and dynamic systems. According to the results, a semantically familiar problem context invited a high number of a priori assumptions on the interdependency of system variables. These assumptions were less likely tested during the knowledge acquisition phase, this proving to be the main barrier to the acquisition of new knowledge. Unsystematic exploration behavior tended to produce non-informative system states that

complicated the extraction of knowledge. A lack of knowledge ultimately led to poor control competency.

Beckmann et al. (2017) confirmed research results by Beckmann and Goode (2014) and demonstrated how a differentiation between complexity and difficulty leads to a better understanding of the cognitive mechanism behind CPS. According to findings from a study with 240 university students, the performance differences observed in the context of the semantic effect were associated with differences in the systematicity of the exploration behavior, and the systematicity of the exploration behavior was reflected in a specific sequence of interventions. They argued that it is only the VOTAT strategy—supplemented with the vary-*none*-at-a-time strategy in the case of noting autonomous changes—that creates informative system state transitions which enable problem solvers to derive knowledge of the causal structure of a complex, dynamic system.

Schoppek and Fischer (2017) also investigated VOTAT and the related “PULSE” strategy (all input variables to zero), which enables the problem solver to observe the eigendynamics of the system in a transfer experiment. They proposed that besides VOTAT and PULSE, other comprehensive knowledge elements and strategies, which contribute to successful CPS, should be investigated.

In a study with 2<sup>nd</sup>- to 4<sup>th</sup>-grade students, Chen and Klahr found little spontaneous development when children interacted with physical objects (in situations similar to that of Piaget’s experiments), while more direct teaching of the control of variables strategy resulted in good effect sizes and older children were able to transfer the knowledge they had acquired (improved control of variable strategy) to remote contexts (Chen and Klahr, 1999). In a more recent study, Kuhn et al. (2008) further extended the scope of studies on scientific thinking, identifying three further aspects beyond the control of variables strategy, including coordinating effects of multiple influences, understanding the epistemological foundations of science and engaging in argumentation. In their experiment with 91 6<sup>th</sup>-grade students, they explored how students were able to estimate the impact of five independent variables simultaneously on a particular phenomenon, and they found that most students considered only one or two variables as possible causes.

## AIMS

In this paper, we explore several research questions on effective and less effective problem-solving strategies used in a complex problem-solving environment and detected by logfile analyses. We use logfile analyses to empirically test the success of different input behavior and strategy usage in CPS tasks within the MicroDYN framework. After constructing a mathematical model based on all theoretically effective strategies, which provide the problem solver with all the information needed to solve the problem, and defining several sub-strategies within the VOTAT strategy based on the amount of effort expended to extract the necessary information, we empirically distinguish different VOTAT and non-VOTAT strategies, which can result in good CPS performance and which go beyond the isolated variation strategy as an effective strategy for rule induction (Vollmeyer

et al., 1996). We highlight the most and least effective VOTAT strategies used in solving MicroDYN problems and empirically investigate the awareness of the strategy used based on the sequence of the sub-strategies used. Based on these results, we conduct latent class analyses to explore and define patterns in qualitatively different VOTAT strategy uses.

We thus intend to answer five research questions:

- RQ1: Does the use of a theoretically effective strategy occur prior to high performance? In other words, does the use of a theoretically effective strategy result in high performance?
- RQ2: Do all VOTAT strategies result in a high CPS performance? What is the most effective VOTAT strategy?
- RQ3: How does awareness of the exploration strategy used influence overall performance on CPS tasks?
- RQ4: What profiles characterize the various problem solvers and explorers?
- RQ5: Do exploration strategy profiles differ across grade levels, which represent different educational stages during compulsory schooling?

## HYPOTHESES

In this study, we investigated qualitatively different classes of students’ exploration behavior in CPS environments. We used latent class analysis (LCA) to study effective and non-effective input behavior and strategy use, especially the principle of isolated variation, across several CPS tasks. We compared the effectiveness of students’ exploration behavior based on the amount of information they extracted with their problem-solving achievement. We posed five separate hypotheses.

Hypothesis 1: We expect that high problem-solving achievement is not closely related to expert exploration behavior.

Vollmeyer et al. (1996) explored the impact of strategy effectiveness on problem-solving performance and reported that effectiveness correlated negatively and weakly to moderately with solution error ( $r = -0.32$  and  $r = -0.54$ ,  $p < 0.05$ ). They reported that “most participants eventually adopted the most systematic strategy, VOTAT, and the more they used it, the better they tended to perform. However, even those using the VOTAT strategy generally did not solve the problem completely” (p. 88). Greiff et al. (2015) confirmed that different exploration behaviors are relevant to CPS and that the number of sub-strategies implemented was related to overall problem-solving achievement.

Hypothesis 2: We expect that students who use the isolated variation strategy in exploring CPS problems have a significantly better overall performance than those who use a theoretically effective, but different strategy.

Sonnleiter et al. (2017) noted that “A more effective 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" (p. 169). According to Wüstenberg et al. (2012), students who consistently apply the principle of isolated variation—the most systematic VOTAT strategy—in CPS environments show better overall CPS performance, compared to those who use different exploration strategies. Kröner et al. (2005) reported a positive correlation between using the principle of isolated variation and the likelihood of solving the overall problem.

Hypothesis 3: We expected that more aware CPS exploration behavior would be more effective than exploration behavior that generally results in extracting all the necessary information from the system to solve the problem, but within which the steps have no logically built structure and no systematicity in time.

Vollmeyer et al. (1996) explored the impact of strategy systematicity on problem-solving performance. They emphasized that "the systematicity of participants' spontaneous hypothesis-testing strategies predicted their success on learning the structure of the biology lab problem space" (p. 88). Vollmeyer and her colleagues restricted systematic strategy users to isolated variation strategy users; this corresponds to our terminology usage of aware isolated variation strategy users.

Hypothesis 4: We expected to find a distinct number of classes with statistically distinguishable profiles of CPS exploration behavior. Specifically, we expected to find classes of proficient, intermediate and low-performing explorers.

Several studies (Osman and Speekenbrink, 2011; Wüstenberg et al., 2012; Greiff et al., 2015) have indicated that there exist quantitative differences between different exploration strategies, which are relevant to a CPS environment. The current study is the first to investigate whether a relatively small number of qualitatively different profiles of students' exploration proficiency can be derived from their behavior detected in a CPS environment in a broad age range.

Hypothesis 5: We expected that more proficient CPS exploration behavior would be more dominant at later grade levels as an indication of cognitive maturation and of increasing abilities to explore CPS environments.

The cognitive development in children between Grades 3 and 12 is immense. According to Piaget's stage theory, they move from concrete operations to formal operations and they will be able to think logically and abstractly. According to Galotti (2011) and Molnár et al. (2013), the ability to solve problems effectively and to make decisions in CPS environments increases in this period of time; Grades 6–8 seem especially crucial for development. Thus, we expect that cognitive maturation will also be reflected in more proficient exploration behavior.

## METHODS

### Participants

The sample was drawn from 3<sup>rd</sup>- to 12<sup>th</sup>-grade students (aged 9–18) in Hungarian primary and secondary schools

( $N = 4,371$ ; **Table 1**). School classes formed the sampling unit. 180 classes from 50 schools in different regions were involved in the study, resulting in a wide-ranging distribution of students' background variables. The proportion of boys and girls was about the same.

### Materials

The MicroDYN approach was employed to develop a measurement device for CPS. CPS tasks within the MicroDYN approach are based on linear structural equations (Funke, 2001), in which up to three input variables and up to three output variables are related (Greiff et al., 2013). Because of the small set of input and output variables, the MicroDYN problems could be understood completely with precise causal analyses (Funke, 2014). The relations are not presented to the problem solver in the scenario. To explore these relations, the problem solver must interact directly with the problem situation by manipulating the input variables (Greiff and Funke, 2010), an action that can influence the output variables (direct effects), and they must use the feedback provided by the computer to acquire and employ new knowledge (Fischer et al., 2012). Output variables can change spontaneously and can consist of internal dynamics, meaning they can change without changing the input variables (indirect effects; Greiff et al., 2013). Both direct and indirect effects can be detected with an adequate problem-solving strategy (Greiff et al., 2012). The interactions between the problem situation and the test taker play an important role, but they can only be identified in a computerized environment based on log data collected during test administration.

In this study, different versions with different levels of item complexity were used (Greiff et al., 2013), which varied by school grade (**Table 2**; six MicroDYN scenarios were administered in total in Grades 3–4; eight in Grade 5; nine in Grades 6–8; and twelve in Grades 9–12); however, we only involved those six tasks where the principle of isolated variation was the optimal exploration strategy. That is, we excluded problems with an external manipulation-independent, internal dynamic effect or multiple dependence effect from the analyses, and there were no delayed or accumulating effects used in the problem environments created. Complexity was defined by the number of input and output variables and the number of relations based on Cognitive Load Theory (Sweller, 1994). "Findings show that

**TABLE 1 |** Composition of samples.

Grade	Sample size	Gender, % female	Mean age (sd)
3	584	–	–
4	679	–	–
5	608	–	–
6	677	49	11.92 (0.53)
7	607	51	12.94 (0.53)
8	942	49	13.89 (0.56)
9	30	48	15.00 (0.59)
10	84	51	16.79 (0.49)
11	102	68	17.02 (0.79)
12	58	64	17.93 (0.57)

**TABLE 2 |** The design of the whole study: the complexity of the systems administered and the structure and anchoring of the tests applied in different grades.

Complexity of the systems (number of input and output variables and connections without internal dynamics)	Presence of autoregressive dependencies	Grade						
		3	4	5	6	7	8	9–12
2-1-2		+	+	+	+	+	+	+
2-2-2		+	+	+	+	+	+	+
2-2-2		+	+	+	+	+	+	+
2-2-2		+	+	+				
3-2-3		+	+	+	+	+	+	+
3-3-3		+	+	+	+	+	+	+
3-3-4								+
3-2-1	+			+	+	+	+	+
3-3-4				+	+	+	+	+
3-2-2	+				+	+	+	+
3-3-3	+				+	+	+	+
3-3-3	+							+
3-3-3	+							+

increases in the number of relations that must be processed in parallel in reasoning tasks consistently lead to increases in task difficulty” (Beckmann and Goode, 2017).

The tasks were designed so that all causal relations could be identified with systematic manipulation of the inputs. The tasks contained up to three input variables and up to three output variables with different fictitious cover stories. The values of the input variables were changed by clicking on a button with a + or – sign or by using a slider connected to the respective variable (see **Figure 1**). The controllers of the input variables range from “–” (value = –2) to “++” (value = +2). The history of the values of the input variables within the same scenario was presented on a graph connected to each input variable. Beyond the input and output variables, each scenario contained a Help, Reset, Apply and Next button. The Reset button set the system back to its original status. The Apply button made it possible to test the effect of the currently set values of the input variables on the output variables, which appeared in the form of a diagram of each output variable. According to the user interface, within the same phase of each of the problem scenarios, the input values remained at the level at which they were set for the previous input until the Reset button was pressed or they were changed manually. The Next button implemented the navigation between the different MicroDYN scenarios and the different phases within a MicroDYN scenario.

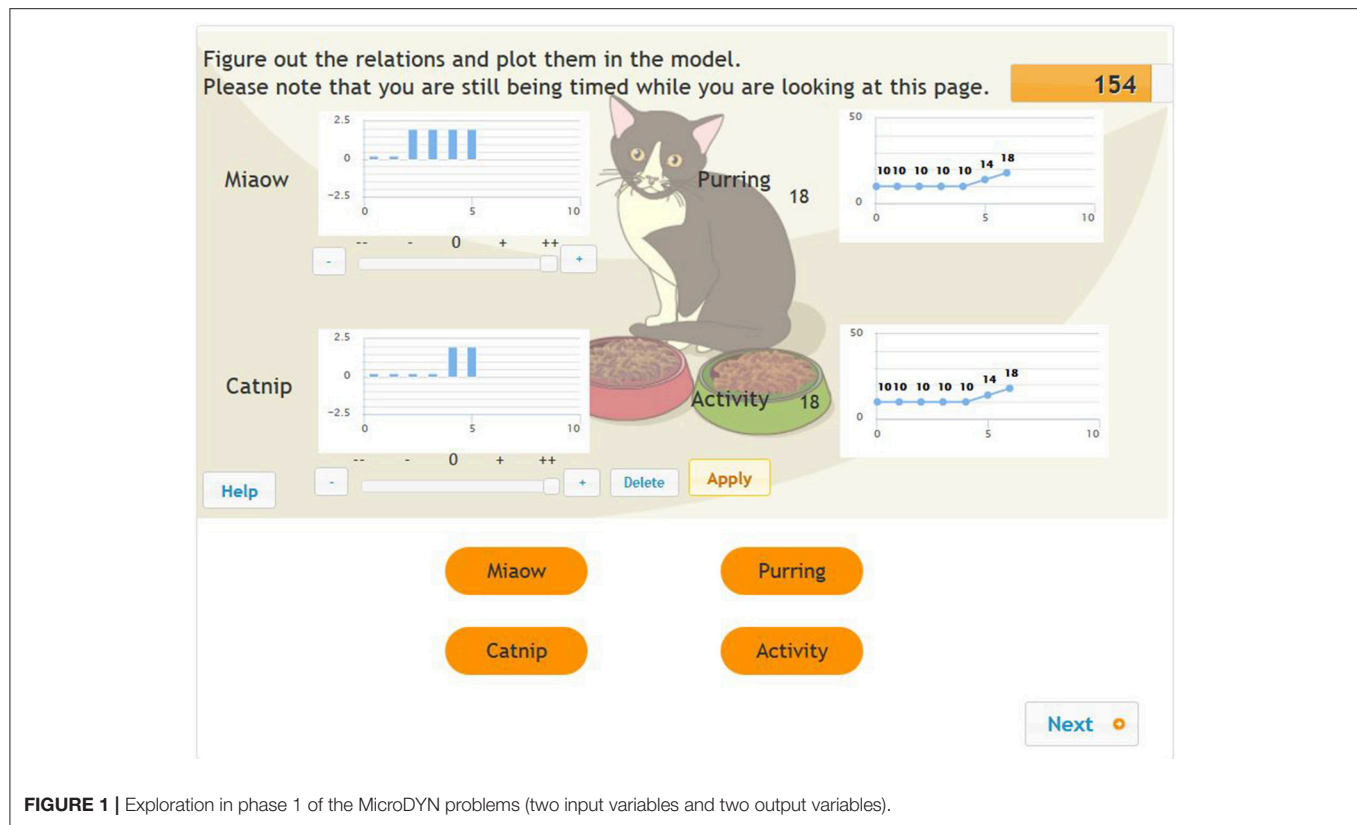
In the knowledge acquisition phase, participants were freely able to change the values of the input variables and attempt as many trials for each MicroDYN scenario as they liked within 180 s. During this 180 s, they had to draw the concept map (or causal diagram; Beckmann et al., 2017); that is, they had to draw the arrows between the variables presented on the concept map under the MicroDYN scenario on screen. In the knowledge application phase, students had to check their respective system using the right concept map presented on screen by reaching the given target values within a given time frame (90 s) in no more than four trials, that is, with a maximum

of four clicks on the Apply button. This applied equally to all participants.

## Procedures

All of the CPS problems were administered online via the eDia platform. At the beginning, participants were provided with instructions about the usage of the user interface, including a warm-up task. Subsequently, participants had to explore, describe and operate unfamiliar systems. The assessment took place in the schools’ ICT labs using the available school infrastructure. The whole CPS test took approximately 45 min to complete. Testing sessions were supervised by teachers who had been thoroughly trained in test administration. Students’ problem-solving performance in the knowledge acquisition and application phases was automatically scored as CPS performance indicators; thus, problem solvers received immediate performance feedback at the end of the testing session. We split the sample into three age groups, whose achievement differed significantly (Grades 3–5,  $N = 1,871$ ; Grades 6–7,  $N = 1,284$ ; Grades 8–12,  $N = 1,216$ ;  $F = 122.56$ ,  $p < 0.001$ ;  $t_{\text{level}_1_2} = -6.22$ ,  $p < 0.001$ ;  $t_{\text{level}_2_3} = -8.92$ ,  $p < 0.001$ ). This grouping corresponds to the changes in the developmental curve relevant to complex problem solving. The most intensive development takes place in Grades 6–7 (see Molnár et al., 2013). Measurement invariance, that is, the issue of structural stability, has been demonstrated with regard to complex problem solving in the MicroDYN approach already (e.g., Greiff et al., 2013) and was confirmed in the present study (**Table 3**). Between group differences can be interpreted as true and not as psychometric differences in latent ability. The comparisons across grade levels are valid.

The latent class analysis (Collins and Lanza, 2010) employed in this study seeks students whose problem-solving strategies show similar patterns. It is a probabilistic or model-based technique, which is a variant of the traditional cluster analysis (Tein et al., 2013). The indicator variables observed were re-coded strategy scores. Robust maximum likelihood estimation



**FIGURE 1** | Exploration in phase 1 of the MicroDYN problems (two input variables and two output variables).

**TABLE 3** | Goodness of fit indices for measurement invariance of MicroDYN problems.

Model	$\chi^2$	df	$\Delta\chi^2$	$\Delta df$	p	CFI	TLI	RMSEA
Configural invariance	119.71	42	—	—	—	0.980	0.987	0.039
Strong factorial invariance	126.33	45	7.37	3	>0.05	0.986	0.980	0.038
Strict factorial invariance	145.49	52	15.02	8	>0.05	0.980	0.976	0.042

$\chi^2$  and df were estimated by the weighted least squares mean and variance adjusted estimator (WLSMV).  $\Delta\chi^2$  and  $\Delta df$  were estimated by the Difference Test procedure in MPlus. Chi-square differences between models cannot be compared by subtracting  $\chi^2$ s and dfs if WLSMV estimators are used. CFI, comparative fit index; TLI, Tucker Lewis index; RMSEA, root mean square error of approximation.

was used and two to seven cluster solutions were examined. The process of latent class analysis is similar to that of cluster analysis. Information theory methods, likelihood ratio statistical test methods and entropy-based criteria were used in reducing the number of latent classes. As a measure of the relative model fit, AIC (Akaike Information Criterion), which considers the number of model parameters, and BIC (Bayesian Information Criterion), which considers the number of parameters and the number of observations, are the two original and most commonly used information theory methods for model selection. The adjusted Bayesian Information Criterion (aBIC) is the sample size-adjusted BIC. Lower values indicated a better model fit for each criterion (see Dziak et al., 2012). Entropy represents the precision of the classification for individual cases. MPlus reports the relative entropy index of the model, which is a re-scaled version of entropy on a [0,1] scale. Values near one, indicating high certainty in classification, and values near zero, indicating low certainty, both point to a low level of homogeneity of the

clusters. Finally, the Lo–Mendell–Rubin Adjusted Likelihood Ratio Test (Lo et al., 2001) was employed to compare the model containing  $n$  latent classes with that containing  $n-1$  latent classes. A significant  $p$ -value ( $p < 0.05$ ) indicates that the  $n-1$  model is rejected in favor of a model with  $n$  classes, as it fits better than the previous one (Muthén and Muthén, 2012).

## Scoring

As previous research has found (Greiff et al., 2013), achievement in the first and second phases of the problem-solving process can be directly linked to the concept of knowledge acquisition (representation) and knowledge application (generating a solution) and was scored dichotomously. For knowledge acquisition, students' responses were scored as correct ("1") if the connections between the variables were accurately indicated on the concept map (students' drawings fully matched the underlying problem structure); otherwise, the response was scored as incorrect ("0"). For knowledge application, students'

responses were scored as correct (“1”) if students reached the given target values within a given time frame and in no more than four steps, that is, with a maximum of four clicks on the Apply button; otherwise, the response was scored as incorrect (“0”).

We developed a labeling procedure to divide the continuum of the problem-solving process into more scoreable phases and to score students’ activity and behavior in the exploration phase at the beginning of the problem-solving process. For the different analyses and the most effective clustering, we applied a categorization, distinguishing students’ use of the full, basic and minimal input behavior within a single CPS task (detailed description see later). The unit of this labeling process was a trial, a setting of the input variables, which was tested by clicking on the Apply button during the exploration phase of a problem, thus between receiving the problem and clicking on the Next button to reach the second part, the application part of the problem. The sum of these trials, within the same problem environment is called the input behavior. The input behavior was called a strategy if it followed meaningful regularities.

By our definition, the full input behavior model describes what exactly was done throughout the exploration phase and what kinds of trials were employed in the problem-solving process. It consists of all the activities with the sliders and Apply buttons in the order they were executed during the first phase, the exploration phase of the problem-solving process. The basic input behavior is part of the full input behavior model by definition, when the order of the trials attempted was still being taken into account, but it only consists of activities where students were able to acquire new information on the system. This means that the following activities and trials were not included in the basic input behavior model (they were deleted from the full input behavior model to obtain the basic behavior model):

- where the same scenario, the same slider adjustment, was employed earlier within the task (that is, we excluded the role of *ad hoc* control behavior from the analyses),
- where the value (position) of more than one input variable (slider) was changed and where the effect of the input variable on the operation of the system was still theoretically unknown to the problem solver,
- where a new setting or new slider adjustment was employed, though the effect of the input variables used was known from previous settings.
- As the basic input behavior involves timing, that is, the order of the trials used, it is suitable for the analyses with regard to the awareness of the input behavior employed.

Finally, we generated the students’ minimal input behavior model from the full input behavior model. By our definition, the minimal input behavior focuses on those untimed activities (a simple list, without the real order of the trials), where students were able to obtain new information from the system and were able to do so by employing the most effective trials.

Each of the activities in which the students engaged and each of the trials which they used were labeled according to the following labeling system to be able to define students’ full input

behavior in a systematic format (please note that the numerical labels are neither scores nor ordinal or metric information):

- Only one single input variable was manipulated, whose relationship to the output variables was unknown (we considered a relationship unknown if its effect cannot be known from previous settings), while the other variables were set at a neutral value like zero. We labeled this trial +1.
- One single input variable was changed, whose relationship to the output variables was unknown. The others were not at zero, but at a setting used earlier. We labeled this trial +2.
- One single input variable was changed, whose relationship to the output variables was unknown, and the others were not at zero; however, the effect of the other input variable(s) was known from earlier settings. Even so, this combination was not attempted earlier. We labeled this trial +3.
- Everything was maintained in a neutral (zero) position. This trial is especially important for CPS problems with their own internal dynamics. We labeled this +A.
- The value of more than one input variable, whose relationship to the output variables was unknown, was changed at the same time, resulting in no additional information on the system. It was labeled -X.
- The same trial, the slider adjustment, had already been employed earlier within the task, resulting in no additional information on the system. It was labeled -0.
- A new slider adjustment was employed; however, the effect of the manipulated input variables was known from previous settings. This trial offered no additional information on the system and was labeled +0.

Although several input variables were changed by the scenario, it was theoretically possible to count the effect of the input variables on the output variables based on the information from the previous and present settings by using and solving linear equations. It was labeled +4.

An extra code (+5) was employed in the labeling process, but only for the basic input behavior, when the problem solver was able to figure out the structure of the problem based on the information obtained in the last trial used. This labeling has no meaning in the case of the minimal input behavior.

The full, basic and minimal input behavior models as well as the labeling procedure can be employed by analyzing problem solvers’ exploration behavior and strategies for problems that are based on minimal complex systems. The user interface can preserve previous input values, and the values are not reset to zero after each exploration input. According to Fischer et al. (2012), VOTAT strategies are best for identifying causal relations between variables and they maximize the successful strategic behavior in minimal complex systems, such as CPS. By using a VOTAT strategy, the problem solver systematically varies only one input variable, while the others remain unchanged. This way, the effect of the changed variable can be found in the system by monitoring the changes in the output variables. There exist several types of VOTAT strategies based on the different combinations of VOTAT-centered trials +1, +2, and +3. The most obvious systematic strategy is when only one input variable



is different from the neutral level in each trial and all the other input variables are systematically maintained at the neutral level. Thus, the strategy is a combination of so-called +1 trials, where it is employed for every input variable. Known as the isolated variation strategy (Müller et al., 2013), this strategy has been covered extensively in the literature. It must be noted that the isolated variation strategy is not appropriate to detect multiple dependence effects within the MicroDYN approach. We hypothesize that there are more and less successful input behaviors and strategies. We expect that theoretically effective, non-VOTAT strategies do not work as successfully as VOTAT strategies and that the most effective VOTAT strategy will be the isolated variation strategy.

We will illustrate the labeling and coding process and the course of generating a minimal input behavior out of a basic or full input behavior through the following two examples.

**Figure 1** shows an example with two input variables and two output variables. (The word problem reads as follows: “When you get home in the evening, there is a cat lying on your doorstep. It is exhausted and can barely move. You decide to feed it, and a neighbor gives you two kinds of cat food, Miaow and Catnip. Figure out how Miaow and Catnip impact activity and purring.”). The student who mapped the operation of the system as demonstrated in the figure pressed the Apply button six times in all, using the various settings for the Miaow and Catnip input variables.

In mapping the system, the problem solver kept the value of both the input variables at 0 in the first two steps (making no changes to the base values of the input variables), as a result of which the values of the output variables remained unchanged. In steps 3 and 4, he set the value of the Miaow input variable at 2, while the value of the Catnip variable remained at 0 (the bar chart by the name of each variable shows the history of these settings). Even making this change had no effect on the values of the output variables; that is, the values in each graph by the purring and activity variables are constantly horizontal. In steps 5 and 6, the student left the value of the Miaow input variable at 2, but a value of 2 was added to this for the Catnip input variable. As a result, the values of both output variables (purring and activity) began to grow by the same amount. The coding containing all the information (the full input behavior) for this sequence of steps was as follows: +A, −0, +1, −0, +2, −0. The reason for this is since steps 2, 4, and 6 were repetitions of previous combinations, we coded them as −0. Step 3 involved the purest use of a VOTAT strategy [changing the value of one input variable at a time, while keeping the values of the other input values at a neutral level (+1)], while the trial used in step 5 was also a VOTAT strategy. After all, only the value of one input variable changed compared to step 4. This is therefore not the same trial as we described in step 3 (+2). After step 5, all the necessary information was available to the problem solver. The basic input behavior for the same sequence of steps was +A, +1, +2, since the rest of the steps did not lead the problem solver to acquire unknown information. Independently of the time factor, the minimal input behavior in this case was also +A, +1, +2. The test taker was able to access new information on the operation of the system through these

steps. From the point of view of awareness, this +1+2 strategy falls under aware strategy usage, as the +1 and +2 sub-strategies were not applied far apart (excluding the simple repetition of the executed trials next to each other) from each other in time. A good indicator of aware strategy usage is if there is no difference between minimal and basic input behavior.

In the second example (**Figure 2**), we demonstrate the sequence of steps taken in mapping another problem as well as the coding we used. Here the students needed to solve a problem consisting of two input variables and one output variable. The word problem reads as follows: “Your mother has bought two new kinds of fruit drink mix. You want to make yourself a fruit drink with them. Figure out how the green and blue powders impact the sweetness of the drink. Plot your assumptions in the model.” The test taker attempted eight different trials in solving this problem, which were coded as follows: +1, +2, +0, +0, +0, +0, −0, −0. After step 2, the student had access to practically all the information required to plot the causal diagram. (In step 1, the problem solver checked the impact of one scoop of green powder and left the quantity of blue powder at zero. Once mixed, the resultant fruit drink became sweeter. In step 2, the problem solver likewise measured out one scoop of green powder for the drink but also added a scoop of blue powder. The sweetness of the drink changed as much as it had in step 1. After that, the student measured out various quantities of blue and then green powder, and looked at the impact.) The basic input behavior coded from the full input behavior used by the problem solver was +1+2, and the minimal input behavior was +1+1 because the purest VOTAT strategy was used in steps 1 and 6. (Thus, both variables separately confirmed the effects of the blue and the green powder on the sweetness of the drink.) From the point of view of awareness, this +1+1 strategy falls under non-aware strategy usage, as the two applications of the +1 trial occurred far apart from each other in time.

Based on students’ minimal input behavior we executed latent class analyses. We narrowed the focus to the principle of isolated variation, especially to the extent to which this special strategy was employed in the exploration phase as an indicator of students’ ability to proficiently explore the problem environment. We added an extra variable to each of the problems, describing students’ exploration behavior based on the following three categories: (1) no isolated variation at all (e.g., isolated variation was employed for none of the input variables – 0 points); (2) partially isolated variation (e.g., isolated variation was employed for some but not all the input variables – 1 point); and (3) fully isolated variation (e.g., isolated variation was employed for all the input variables – 2 points). Thus, depending on the level of optimal exploration strategy used, all the students received new categorical scores based on their input exploration behavior, one for each of the CPS tasks. Let us return to the example provided in **Figures 1, 2**. In the first example, a partially isolated strategy was applied, since the problem solver only used this strategy to test the effect of the Miaow input variables (in trials 3 and 4). In the second example, a full isolated strategy was applied, as the problem solver used this isolated variation strategy for both the input variables during the exploration phase in the first and sixth trials.

RESULTS

The Reliability of the Test Improved When Scoring Was Based on the Log Data

The reliability of the MicroDYN problems as a measure of knowledge acquisition and knowledge application, the traditional CPS indicators for phases 1 and 2, were acceptable at  $\alpha = 0.72\text{--}0.86$  in all grades (Table 4). After we re-scored the problem solvers' behavior at the beginning of the problem-solving process,

coded the log data and assigned new variables for the effectiveness of strategy usage during the exploration phase of the task for each task and person, the overall reliability of the test scores improved. This phenomenon was noted in all grades and in both coding procedures, when the amount of information obtained was examined (Cronbach's  $\alpha$  ranged from 0.86 to 0.96) and when the level of optimal exploration strategy used was analyzed (Cronbach's  $\alpha$  ranged from 0.83 to 0.98; the answers to the warm-up tasks were excluded from these analyses).

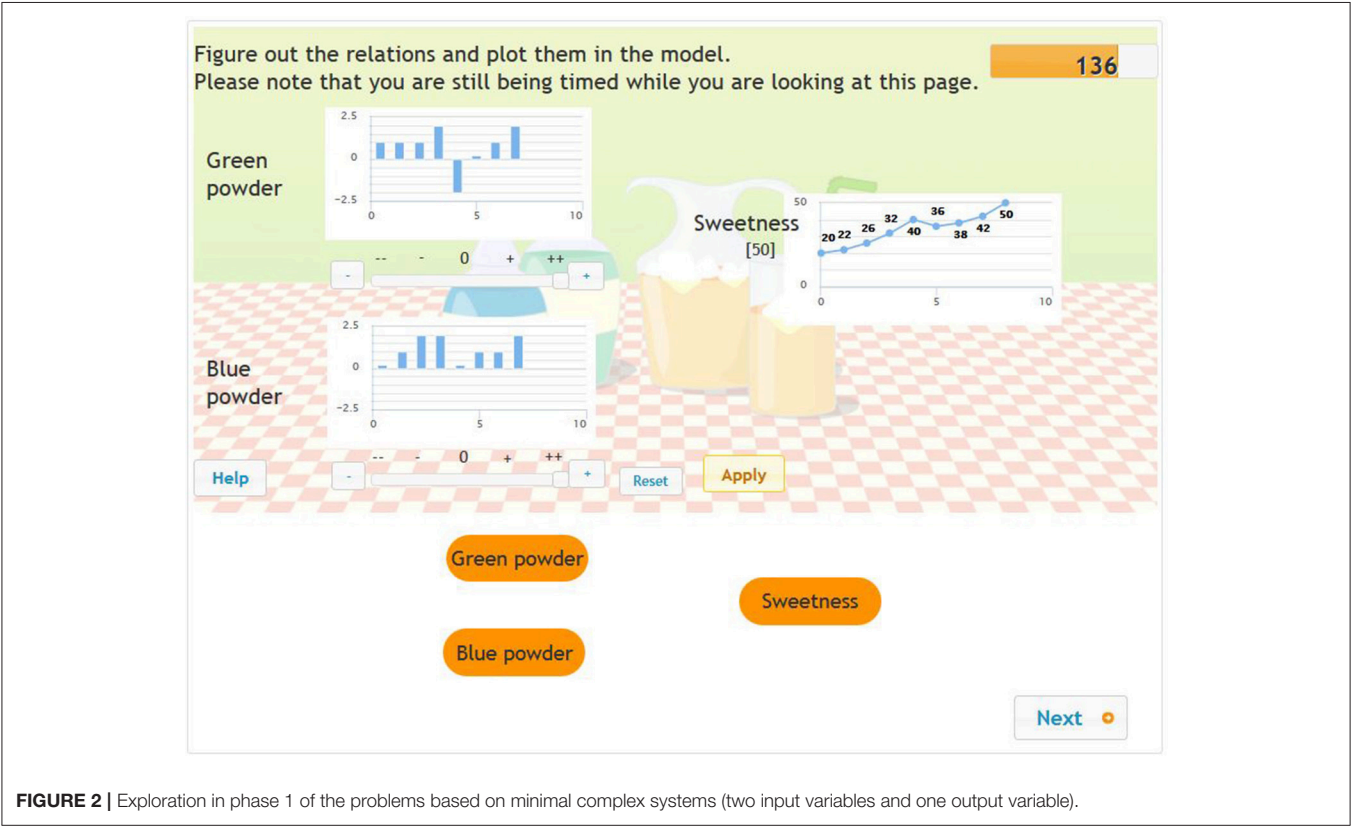


TABLE 4 | Internal consistencies in scoring the MicroDYN problems: analyses based on both traditional CPS indicators and re-coded log data based on student behavior at the beginning of the problem-solving process.

Grade	Reliabilities of the test by traditional scoring (phases 1 and 2)	Reliabilities of the test consisting of the new dichotomously scored variables in terms of the effectiveness of strategy usage at the beginning of the problem-solving process	Reliabilities of the test consisting of traditional scored items and the new dichotomously scored variables describing the effectiveness of strategy usage	Reliabilities of the test consisting of the new categorically scored variables describing the level of isolated variation strategy usage
3	0.83	0.87	0.80	0.83
4	0.77	0.86	0.85	0.86
5	0.78	0.90	0.88	0.90
6	0.72	0.91	0.88	0.93
7	0.74	0.92	0.89	0.94
8	0.80	0.92	0.90	0.95
9	0.83	0.96	0.93	0.97
10	0.85	0.94	0.93	0.96
11	0.86	0.94	0.93	0.98
12	0.83	0.93	0.92	0.97

## Use of a Theoretically Effective Strategy Does Not Result in High Performance (RQ1)

Use of a theoretically effective strategy did not always result in high performance. The percentage of effective strategy use and high CPS performance varied from 20 to 80%, depending on the complexity of the CPS tasks and the age group. The percentage of theoretically effective strategy use in each cohort increased by 20% for age when problems with the same complexity were compared (Table 5) and decreased about 20% for the increasing number of input variables in the problems.

The percentage of theoretically effective strategy use was the same for the less complex problems in Grades 3–5 and for the most complex tasks in Grades 8–12 (58%). More than 80% of these students solved the problem correctly in the first case, but only 60% had the correct solution in the second case. There was a 50% probability of effective and non-effective strategy use for problems with two input and two output variables in Grades 3–5 and for problems with three input and three output variables in Grades 6–7. In Grades 8–12, the use of a theoretically effective strategy was always higher than 50%, independently of the complexity of the problems (with no internal dynamic). The guessing factor, that is, the *ad hoc* optimization (use of a theoretically non-effective strategy with the correct solution) also changed, mostly based on the complexity and position of the tasks in the test. The results confirmed our hypothesis that the use of a theoretically effective strategy does not necessarily represent the correct solution and that the correct solution does not always represent the use of an even theoretically effective problem-solving strategy.

## Not All the VOTAT Strategies Result in High CPS Performance (RQ2)

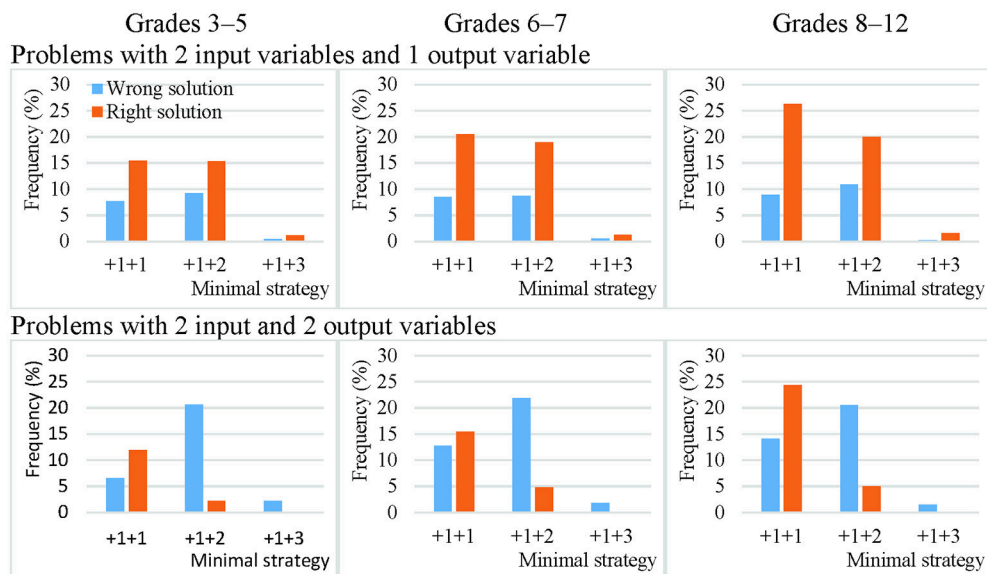
On average, only 15% of the theoretically effective strategy uses involved non-VOTAT strategies. The isolated variation strategy comprised 45% of the VOTAT strategies employed. It was the only theoretically effective strategy which always resulted in the correct solution to the problem with higher probability independently of problem complexity or the grade of the students. The real advantage of this strategy was most remarkable in the case of the third cohort, where an average of 80% of the students who employed this strategy solved the problems correctly (Figures 3, 4).

The second most frequently employed and successful VOTAT strategy was the +1+2 type or the +1+2+2 type, depending on the number of input variables. In the +1+2 type, only one single input variable was manipulated in the first step, while the other variable remained at a neutral value; in the second step, only the other input variable was changed and the first retained the setting used previously. This proved to be relatively successful on problems with a low level of complexity independently of age, but it generally resulted in a good solution with a low level of probability on more complex problems.

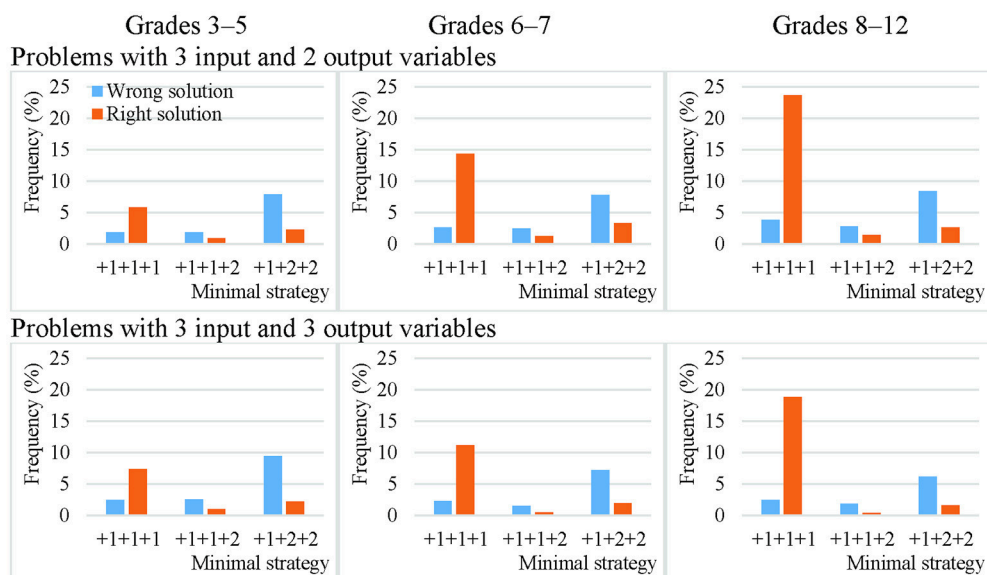
VOTAT strategies of the +1+3 type (in the case of two input variables) and of the +1+1+2 type (in the case of three input variables) were employed even less frequently and with a lower level of efficacy than all the other VOTAT strategies (+1+1+3, +1+2+1, +1+2+2, +1+2+3, +1+3+1, +1+3+2 and +1+3+3 in the case of three input variables) and theoretically effective, non-VOTAT strategies (e.g., +4 in the case of two input variables or +1+4, +4+2 and +4+3 in the case of three input variables). In the following, we provide an example

**TABLE 5 |** Percentage of theoretically effective and non-effective strategy use and high CPS performance.

Complexity of problem (number of input and output variables and connections)	Frequency (%)					
	Theoretically effective strategy use			Theoretically non-effective strategy use		
	Low achievement (%; in proportion to whole sample)	High achievement (%; in proportion to whole sample)	Sum	Low achievement (%; in proportion to whole sample)	High achievement (%; in proportion to whole sample)	Sum
<b>GRADES 3–5</b>						
2-1 (2)	19.9 (11.6)	80.1 (46.6)	58.2	28.2 (11.8)	71.8 (30.0)	41.8
2-2 (2)	81.5 (39.8)	18.5 (9.0)	50.2	97.2 (46.8)	2.8 (1.4)	49.8
3-2 (3)	65.9 (21.5)	34.1 (11.1)	32.6	89.3 (60.2)	10.7 (7.2)	67.4
3-3 (3)	60.2 (21.9)	39.8 (14.5)	36.4	77.1 (49.0)	22.9 (14.6)	63.6
<b>GRADES 6–7</b>						
2-1 (2)	28.3 (18.7)	71.6 (47.2)	65.9	26.9 (9.2)	73.1 (24.9)	34.1
2-2 (2)	72.4 (47.0)	27.5 (18.0)	59.0	98.2 (34.4)	1.8 (0.6)	41.0
3-2 (3)	50.8 (22.9)	49.2 (22.2)	45.0	85.9 (47.2)	14.1 (7.8)	54.9
3-3 (3)	52.6 (25.7)	47.4 (23.2)	49.0	77.3 (39.5)	22.7 (11.6)	51.0
<b>GRADES 8–12</b>						
2-1 (2)	28.7 (21.9)	71.3 (54.5)	76.4	25.5 (6.0)	74.5 (17.6)	23.6
2-2 (2)	59.4 (43.2)	40.6 (29.5)	72.7	98.2 (26.8)	1.8 (0.5)	27.3
3-2 (3)	42.0 (22.8)	58.0 (31.4)	54.2	81.9 (37.5)	18.1 (8.3)	45.8
3-3 (3)	39.4 (22.8)	60.6 (35.2)	58.0	74.1 (31.2)	25.8 (10.9)	42.0



**FIGURE 3 |** Efficacy of the most frequently employed VOTAT strategies on problems with two input variables and one or two output variables in Grades 3–5, 6–7, and 8–12.



**FIGURE 4 |** Efficacy of the most frequently employed VOTAT strategies on problems with three input variables and one or two output variables in Grades 3–5, 6–7, and 8–12.

of the +4+2 type, where the MicroDyn problem has three input variables (A, B, and C) and three output variables. In the first trial, the problem solver set the input variables to the following values: 0 (for variable A), 1 (for variable B), and 1 (for variable C); that is, he or she changed two input variables at the same time. In the second trial, he or she changed the value of two input variables at the same time again and applied the following setting: 0 (for variable A), -2 (for variable B), and -1 (for variable C). In the third trial, he set variable A to 1, and left variables B and C unchanged. That is, the problem solver's input behavior can

be described with the following trials: -X +4 +2. Based on this strategy, it was possible to map the relationships between the input and output variables without using any VOTAT strategy in the exploration phase.

### Aware Explorers Perform Significantly Higher on the CPS Tasks (RQ3)

We compared the achievement of the aware, isolated strategy users with that of the non-aware explorers (Table 6). The percentage of high achievers among the non-aware explorers



**TABLE 6 |** Percentage of high achievers among aware and non-aware explorers by grade and problem complexity.

Complexity of problem	Grades 3–5				Grades 6–7				Grades 8–12			
	Percentage of high achievers among non-aware explorers	Percentage of high achievers among aware explorers using isolated variation strategy	Differences	Percentage of high achievers among aware explorers	Percentage of high achievers among aware explorers using isolated variation strategy	Differences	Percentage of high achievers among non-aware explorers	Percentage of high achievers among aware explorers using isolated variation strategy	Differences	Percentage of high achievers among non-aware explorers	Percentage of high achievers among aware explorers using isolated variation strategy	Differences
2-1	57.99	74.51	16.52	70.72	78.98	8.26	72.29	83.43	11.14			
2-2	3.59	62.94	59.35	2.87	70.55	67.68	3.52	80.92	77.4			
3-2	12.04	77.88	65.84	17.91	84.51	66.6	19.28	88.29	69.01			
3-3	24.68	79.47	54.79	23.17	88.61	65.44	26.84	91.28	64.44			

seemed to be almost independent of age, but strongly influenced by the complexity of the problem and the learning effect we noted in the testing procedure (see RQ5). Results for problems with two input variables and one output variable confirmed our previous results, which showed that the probability of providing the correct solution is very high even without aware use of a theoretically effective strategy (60–70%). With more complex problems, the difference between the percentages of aware and non-aware explorers was huge. Generally, 85% of the non-aware explorers failed on the problems, while at least 80% of the aware, isolated strategy users were able to solve the problems correctly.

### Six Qualitatively Different Explorer Class Profiles Can Be Distinguished at the End of the Elementary Level and Five at the End of the Secondary Level (RQ4 and RQ5)

In all three cohorts, each of the information theory criteria used (AIC, BIC, and aBIC) indicated a continuous decrease in an increasing number of latent classes. The likelihood ratio statistical test (Lo–Mendell Rubin Adjusted Likelihood Ratio Test) showed the best model fit in Grades 3–5 for the 4-class model, in Grades 6–7 for the 6-class model and in Grades 8–12 for the 5-class model. The entropy-based criterion reached the maximum values for the 2- and 3-class solutions, but it was also high for the best-fitting models based on the information theory and likelihood ratio criteria. Thus, the entropy index for the 4-class model showed that 80% of the 3<sup>rd</sup>- to 5<sup>th</sup>-graders, 82% of the 6<sup>th</sup>- to 7<sup>th</sup>-graders and 85% of the 8<sup>th</sup>- to 12<sup>th</sup>-graders were accurately categorized based on their class membership (Table 7).

We distinguished four latent classes in the lower grades based on the exploration strategy employed and the level of isolated variation strategy used (Table 8): 40.5% of the students proved to be non-performing explorers on the basis of their strategic patterns in the CPS environments. They did not use any isolated or partially isolated variation at all; 23.6% of the students were among the low-performing explorers who only rarely employed a fully or partially isolated variation strategy (with 0–20% probability on the less complex problems and 0–5% probability on the more complex problems). 24.7% of the 3<sup>rd</sup>- to 5<sup>th</sup>-graders were categorized as slow learners who were intermediate performers with regard to the efficiency of the exploration strategy they used on the easiest problems with a slow learning effect, but low-performing explorers on the complex ones. In addition, 11.1% of the students proved to be proficient explorers, who used the isolated or partially isolated variation strategy with 80–100% probability on all the proposed CPS problems.

In Grades 6–7, in which achievement proved to be significantly higher on average, 10% fewer students were observed in each of the first two classes (non-performing explorers and low-performing explorers). The percentage of intermediate explorers remained almost the same (26%), and we noted two more classes with the analyses: the class of rapid learners (4.4%) and that of slow learners, who are almost proficient explorers on the easiest problems, employing the fully or partially isolated variation strategy with 60–80% probability,

**TABLE 7 |** Information theory, likelihood ratio and entropy-based fit indices for latent class analyses.

Number of latent classes	AIC	BIC	aBIC	Entropy	L–M–R test	<i>p</i>
<b>GRADES 3–5</b>						
2	13,383	13,512	13,433	0.854	2,301	0.001
3	12,687	12,883	12,763	0.870	714	0.001
4	12,563	12,826	12,664	0.800	149	0.006
5	12,448	12,778	12,574	0.766	139	0.051
6	12,362	12,758	12,514	0.782	110	0.100
<b>GRADES 6–7</b>						
2	13,383	13,512	13,433	0.854	2,301	0.001
3	12,751	12,947	12,826	0.873	1,068	0.001
4	12,576	12,840	12,678	0.819	198	0.001
5	12,497	12,827	12,624	0.814	104	0.004
6	12,427	12,824	12,580	0.823	95	0.022
7	12,402	12,866	12,580	0.828	50	0.498
<b>GRADES 8–12</b>						
2	8,232	8,319	8,265	0.941	2,197	0.001
3	7,718	7,850	7,768	0.856	524	0.001
4	7,690	7,869	7,757	0.829	44	0.002
5	7,686	7,911	7,771	0.853	21	0.003
6	7,705	7,976	7,807	0.770	4	0.561

AIC, Akaike Information Criterion; BIC, Bayesian Information Criterion; aBIC, adjusted Bayesian Information Criterion; L–M–R test, Lo–Mendell–Rubin Adjusted Likelihood Ratio Test. The best fitting model solution is in *italics*.

but low-performing explorers on the complex ones (10.3%). The frequency of proficient strategy users was also increased (to 14.2%) compared to students in the lower grades. Finally, there was almost no change detected in the low-performing explorers' classes in Grades 8–12. We did not detect anyone in the class of intermediate explorers; they must have developed further and become (1) rapid learners (7.7%), (2) slow learners with almost high achievement with regard to the exploration strategy they used on the easiest problems, but low achievers on the complex ones (17.6%), or (3) proficient strategy users (26.3%), whose achievement was high both on the simplest and the most complex problems.

Based on these results, the percentage of non- and low explorers, who have very low exploration skills and do not learn during testing, decreased from almost 65 to 50% between the lower and higher primary school levels and then remained constant at the secondary level. There was a slight increase in respect of the percentage of students among the rapid learners. The students in that group used the fully or partially isolated strategy at very low levels at the beginning of the test, but they learned very quickly and detected these effective exploration strategies; thus, by the end of the test, their proficiency level with regard to exploration was equal to the top performers' achievement. However, we were unable to detect the class of rapid learners among 3<sup>rd</sup>- to 5<sup>th</sup>-graders.

Generally, students' level of exploration expertise with regard to fully and partially isolated variation improved significantly with age ( $F = 70.376$ ,  $p < 0.001$ ). According to our expectations

based on the achievement differences among students in Grades 3–5, 6–8 and 9–12, there were also significant differences in the level of expertise in fully or partially isolated strategy use during problem exploration between 3<sup>rd</sup>- to 5<sup>th</sup>- and 6<sup>th</sup>- to 7<sup>th</sup>-grade students ( $t = -6.833$ ,  $p < 0.001$ ,  $d = 0.03$ ) and between 6<sup>th</sup>- to 7<sup>th</sup>- and 8<sup>th</sup>- to 12<sup>th</sup>-grade students ( $t = -6.993$ ,  $p < 0.001$ ,  $d = 0.03$ ).

## DISCUSSION

In this study, we examined 3<sup>rd</sup>- to 12<sup>th</sup>-grade (aged 9–18) students' problem-solving behavior by means of a logfile analysis to identify qualitatively different exploration strategies. Students' activity in the first phase of the problem-solving process was coded according to a mathematical model that was developed based on strategy effectiveness and then clustered for comparison. Reliability analyses of students' strategy use indicated that strategies used in the knowledge acquisition phase described students' development (ability level) better than traditional quantitative psychometric indicators, including the goodness of the model. The high reliability indices indicate that there are untapped possibilities in analyzing log data. Our analyses of logfiles extracted from a simulation-based assessment of problem solving have expanded the scope of previous studies and made it possible to identify a central component of children's scientific reasoning: the way students understand how scientific experiments can be designed and how causal relationships can be explored by systematically changing the values of (independent) variables and observing their impact on other (target) variables.

In this way, we have introduced a new labeling and scoring method that can be employed in addition to the two scores that have already been used in previous studies. We have found that using this scoring method (based on student strategy use) improves the reliability of the test. Further studies are needed to examine the validity of the scale based on this method and to determine what this scale really measures. We may assume that the general idea of varying the values of the independent variables and connecting them to the resultant changes in the target variable is the essence of scientific reasoning and that the systematic manipulation of variables is related to combinatorial reasoning, while summarizing one's observations and plotting a model is linked to rule induction. Such further studies have to place CPS testing in the context of other cognitive tests and may contribute to efforts to determine the place of CPS in a system of cognitive abilities (see e.g., Wüstenberg et al., 2012).

We have found that the use of a theoretically effective strategy does not always result in high performance. This is not surprising, and it confirms research results by de Jong and van Joolingen (1998), who argue that learners often have trouble interpreting data. As we observed earlier, using a systematic strategy requires combinatorial thinking, while drawing a conclusion from one's observations requires rule induction (inductive reasoning). Students showing systematic strategies but failing to solve the problem may possess combinatorial skills but lack the necessary level of inductive reasoning. It is more difficult to find an explanation for the other direction of discrepancy, when students actually solve the problem without an effective (complete)

**TABLE 8 |** Relative frequencies and average latent class probabilities across grade levels 3–5, 6–7, and 8–12.

Profiles	Grades 3–5		Grades 6–7		Grades 8–12	
	Frequency	Average latent class probabilities	Frequency	Average latent class probabilities	Frequency	Average latent class probabilities
Non-performers	40.5	0.94	30.9	0.92	32.2	0.92
Low performers	23.6	0.86	14.0	0.84	16.2	0.87
Intermediate performers on easiest problems, but low performers on complex ones with a very slow learning effect	24.7	0.82	26.2	0.86	–	–
Rapid learners	–	–	4.4	0.86	7.7	0.96
Almost high performers on easiest problems, but low performers on complex ones with a slow learning effect	–	–	10.3	0.82	17.6	0.79
Proficient strategy users	11.1	0.97	14.2	0.96	26.3	0.97

strategy. Thus, solving the problem does not require the use of a strategy which provides the problem solver with sufficient information about the problem environment to be able to form the correct solution. This finding is similar to results from previous research (e.g., Vollmeyer et al., 1996; Greiff et al., 2015). Goode and Beckmann (2010) reported two qualitatively different, but equally effective approaches: knowledge- based and *ad hoc* control.

In the present study, the contents of the problems were not based on real knowledge, and the causal relationships between the variables were artificial. Content knowledge was therefore no help to the students in filling the gap between the insufficient information acquired from interaction and the successful solution to the problem. We may assume that students guessed intuitively in such a case. Further studies may ascertain how students guess in such situations.

The percentage of success is influenced by the complexity of the CPS tasks, the type of theoretically effective strategy used, the age group and, finally, the degree to which the strategy was consciously employed.

The most frequently employed effective strategies fell within the class of VOTAT strategies. Almost half the VOTAT strategies were of the isolated variation strategy type, which resulted with higher probability in the correct solution independently of the complexity of the problem or the grade of the students. As noted earlier, not all the VOTAT strategies resulted in high CPS performance; moreover, all the other VOTAT strategies proved to be significantly less successful. Some of them worked with relative success on problems with a low level of complexity, but failed with a high level of probability on more complex problems independently of age group. Generally, the advantage of the isolated variation strategy (Wüstenberg et al., 2014) compared to the other VOTAT and non-VOTAT, theoretically effective strategies is clearly evident from the outcome. The use of the isolated variation strategy, where students examined the effect of the input variables on the output variables independently, resulted in a good solution with the highest probability and proved to be the most effective VOTAT strategy independently of student age or problem complexity.

Besides the type of strategy used, awareness also played an influential role. Aware VOTAT strategy users proved to be the most successful explorers. They were followed in effectiveness by non-aware VOTAT strategy users and theoretically effective, but non-VOTAT strategy users. They managed to represent the information that they had obtained from the system more effectively and made good decisions in the problem-solving process compared to their peers.

We noted both qualitative and quantitative changes of problem-solving behavior in the age range under examination. Using latent class analyses, we identified six qualitatively different class profiles during compulsory schooling. (1) Non-performing and (2) low-performing students who usually employed no fully or partially isolated variation strategy at all or, if so, then rarely. They basically demonstrated unsystematic exploration behavior. (3) Proficient strategy users who consistently employed optimal exploration strategies from the very first problem as well as the isolated variation strategy and the partially isolated variation, but only seldom. They must have more elaborated schemas available. (4) Slow learners who are intermediate performers on the easiest problems, but low performers on the complex ones or (5) high performers on the easiest problems, but low performers on the complex ones. Most members of this group managed to employ the principle of isolated or partially isolated variation and had an understanding of it, but they were only able to use it on the easiest task and then showed a rapid decline on the more complex CPS problems. They might have been cognitively overloaded by the increasingly difficult problem-solving environments they faced. (6) Rapid learners, a very small group from an educational point of view. These students started out as non-performers in their exploration behavior on the first CPS tasks, showed a rapid learning curve afterwards and began to use the partially isolated variation strategy increasingly and then the fully isolated variation strategy. By the end of the test, they reached the same high level of exploration behavior as the proficient explorers. We observed no so-called intermediate strategy users, i.e., those who used the partially isolated variation strategy almost exclusively on the test. As we expected, class membership increased significantly in the more proficient classes at the higher grade levels due to the

effects of cognitive maturation and schooling, but this did not change noticeably in the two lowest-level classes.

Limitations of the study include the low sample size for secondary school students; further, repetition is required for validation. The generalizability of the results is also limited by the effects of semantic embedding (i.e., cover stories and variable labels), that is, the usage of different fictitious cover stories “with the intention of minimizing the uncontrollable effects of prior knowledge, beliefs or suppositions” (Beckmann and Goode, 2017). An assumption triggered by semantic contexts has an impact on exploration behavior (e.g., the range of interventions, or strategies employed by the problem solver; Beckmann and Goode, 2014), that is, how the problem solver interacts with the system. Limitations also include the characteristics of the interface used. In our view, analyses with regard to VOTAT strategies are only meaningful in systems with an interface where inputs do not automatically reset to zero from one input to the next (Beckmann and Goode, 2017). That is, we excluded problem environments from the study where the inputs automatically reset to zero from one input to the next. A further limitation of the generalizability of the results is that we have omitted problems with autonomic changes from the analyses.

The main reason why we have excluded systems that contain autoregressive dependencies from the analyses is that different strategy usage is required on problems which also involve the use of trial +A (according to our coding of sub-strategies), which is not among the effective sub-strategies for problems without autonomic changes. Analyses of students' behavior on problems with autonomic changes will form part of further studies, as well as a refinement of the definition of what makes a problem complex and difficult. We plan to adapt the Person, Task and Situation framework published by Beckmann and Goode (2017). The role of *ad hoc* control behavior was excluded from the analyses; further studies are required to ascertain the importance of the repetitive control behavior. Another limitation of the study

could be the interpretation of the differences across age group clusters as indicators of development and not as a lack of stability of the model employed.

These results shed new light on and provide a new interpretation of previous analyses of complex problem solving in the MicroDYN approach. They also highlight the importance of explicit enhancement of problem-solving skills and problem-solving strategies as a tool for applying knowledge in a new context during school lessons.

## ETHICS STATEMENT

Ethical approval was not required for this study based on national and institutional guidelines. The assessments which provided data for this study were integrated parts of the educational processes of the participating schools. The coding system for the online platform masked students' identity; the data cannot be connected to the students. The results from the no-stakes diagnostic assessments were disclosed only to the participating students (as immediate feedback) and to their teachers. Because of the anonymity and no-stakes testing design of the assessment process, it was not required or possible to request and obtain written informed parental consent from the participants.

## AUTHOR CONTRIBUTIONS

Both the authors, GM and BC, certify that they have participated sufficiently in the work to take responsibility for the content, including participation in the concept, design and analysis as well as the writing and final approval of the manuscript. Each author agrees to be accountable for all aspects of the work.

## FUNDING

This study was funded by OTKA K115497.

## REFERENCES

- OECD (2014). *PISA 2012 Results: Creative Problem Solving. Students' Skills in Tackling Real-Life Problems (Volume V)*. Paris: OECD.
- Beckmann, J. F., and Goode, N. (2014). The benefit of being naïve and knowing it? The unfavourable impact of perceived context familiarity on learning in complex problem solving tasks. *Instructional Sci.* 42, 271–290. doi: 10.1007/s11251-013-9280-7
- Beckmann, J. F., Birney, D. P. and Goode, N. (2017). Beyond psychometrics: the difference between difficult problem solving and complex problem solving. *Front. Psychol.* 8:1739. doi: 10.3389/fpsyg.2017.01739
- Beckmann, J. F., and Goode, N. (2017). Missing the wood for the wrong trees: On the difficulty of defining the complexity of complex problem solving scenarios. *J. Intell.* 5:2. doi: 10.3390/jintelligence5020015
- Chen, Z., and Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child Dev.* 70, 1098–1120. doi: 10.1111/1467-8624.00081
- Collins, L. M., and Lanza, S. T. (2010). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. New York, NY: Wiley.
- Csapó, B., Ainley, J., Bennett, R., Latour, T., and Law, N. (2012). “Technological issues of computer-based assessment of 21st century skills,” in *Assessment and Teaching of 21st Century Skills*, eds P. Griffin, B. McGaw, and E. Care (New York, NY: Springer), 143–230.
- Csapó, B., and Funke, J. (eds.). (2017). *The Nature of Problem Solving. Using Research to Inspire 21st Century Learning*. Paris: OECD.
- Csapó, B., and Molnár, G. (2017). Potential for assessing dynamic problem-solving at the beginning of higher education studies. *Front. Psychol.* 8:2022. doi: 10.3389/fpsyg.2017.02022
- Csapó, B., Molnár, G., and Nagy, J. (2014). Computer-based assessment of school-readiness and reasoning skills. *J. Educ. Psychol.* 106, 639–650. doi: 10.1037/a0035756
- de Jong, T., and van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Rev. Educ. Res.* 68, 179–201. doi: 10.3102/00346543068002179
- Demetriou, A., Efklides, A., and Platsidou, M. (1993). The architecture and dynamics of developing mind: experiential structuralism as a frame for unifying cognitive developmental theories. *Monogr. Soc. Res. Child Dev.* 58, 1–205. doi: 10.2307/1166053
- Dörner, D., and Funke, J. (2017). Complex problem solving: what it is and what it is not. *Front. Psychol.* 8:1153. doi: 10.3389/fpsyg.2017.01153



- Dziak, J. J., Coffman, D. L., Lanza, S. T., and Li, R. (2012). *Sensitivity and Specificity of Information Criteria*. The Pennsylvania State University: The Methodology Center and Department of Statistics. Available online at: <https://methodology.psu.edu/media/techreports/12-119.pdf>
- Fischer, A., Greiff, S., and Funke, J. (2012). The process of solving complex problems. *J. Probl. Solving*, 4, 19–42. doi: 10.7771/1932-6246.1118
- Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Think. Reason*, 7, 69–89. doi: 10.1080/13546780042000046
- Funke, J. (2010). Complex problem solving: a case for complex cognition? *Cogn. Process*, 11, 133–142. doi: 10.1007/s10339-009-0345-0
- Funke, J. (2014). Analysis of minimal complex systems and complex problem solving require different forms of causal cognition. *Front. Psychol.* 5:739. doi: 10.3389/fpsyg.2014.00739
- Galotti, K. M. (2011). *Cognitive Development*. Thousand Oaks, CA: SAGE.
- Gardner, P. H., and Berry, D. C. (1995). The effect of different forms of advice on the control of a simulated complex system. *Appl. Cogn. Psychol.* 9:7. doi: 10.1002/acp.2350090706
- Goode, N., and Beckmann, J. F. (2010). You need to know: there is a causal relationship between structural knowledge and control performance in complex problem solving tasks. *Intelligence* 38, 345–352. doi: 10.1016/j.intell.2010.01.001
- Greiff, S., and Funke, J. (2010). Systematische Erforschung komplexer Problemlösefähigkeit anhand minimal komplexer Systeme. *Zeitschrift für Pädagogik* 56, 216–227.
- Greiff, S., and Funke, J. (2017). Interactive problem solving: exploring the potential of minimal complex systems,” in *The Nature of Problem Solving. Using Research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (Paris: OECD), 93–105.
- Greiff, S., Wüstenberg, S., and Avvisati, F. (2015). Computer-generated log-file analyses as a window into students’ minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers Educ.* 91, 92–105. doi: 10.1016/j.compedu.2015.10.018
- Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, H., Graesser, A. C., et al. (2014). Domain-general problem solving skills and education in the 21st century. *Educ. Res. Rev.* 13, 74–83. doi: 10.1016/j.edurev.2014.10.002
- Greiff, S., Wüstenberg, S., and Funke, J. (2012). Dynamic problem solving: a new assessment perspective. *Appl. Psychol. Meas.* 36, 189–213. doi: 10.1177/0146621612439620
- Greiff, S., Wüstenberg, S., Goetz, T., Vainikainen, M.-P., Hautamäki, J., and Bornstein, M. H. (2015). A longitudinal study of higher-order thinking skills: working memory and fluid reasoning in childhood enhance complex problem solving in adolescence. *Front. Psychol.* 6:1060. doi: 10.3389/fpsyg.2015.01060
- Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., and Csapó, B. (2013). Complex problem solving in educational contexts – Something beyond g: Concept, assessment, measurement invariance, and construct validity. *J. Educ. Psychol.* 105, 364–379. doi: 10.1037/a0031856
- Griffin, P., McGaw, B., and Care, E. (2012). *Assessment and Teaching of 21st Century Skills*. Dordrecht: Springer.
- Inhelder, B., and Piaget, J. (1958). *The Growth of Logical Thinking from Childhood to Adolescence*. New York, NY: Basic Books.
- Klahr, D., and Dunbar, K. (1988). Dual space search during scientific reasoning. *Cogn. Sci.* 12, 1–48. doi: 10.1207/s15516709cog1201\_1
- Klahr, D., Triona, L. M., and Williams, C. (2007). Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. *J. Res. Sci. Teaching* 44, 183–203. doi: 10.1002/tea.20152
- Kröner, S., Plass, J. L., and Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence* 33, 347–368. doi: 10.1016/j.intell.2005.03.002
- Kuhn, D., Iordanou, K., Pease, M., and Wirkala, C. (2008). Beyond control of variables: what needs to develop to achieve skilled scientific thinking? *Cogn. Dev.* 23, 435–451. doi: 10.1016/j.cogdev.2008.09.006
- Lo, Y., Mendell, N. R., and Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika* 88, 767–778. doi: 10.1093/biomet/88.3.767
- Lotz, C., Scherer, R., Greiff, S., and Sparfeldt, J. R. (2017). Intelligence in action – Effective strategic behaviors while solving complex problems. *Intelligence* 64, 98–112. doi: 10.1016/j.intell.2017.08.002
- Mayer, R. E., and Wittrock, M. C. (1996). “Problem-solving transfer” in *Handbook of Educational Psychology*, eds D. C. Berliner and R. C. Calfee (New York, NY; London: Routledge), 47–62.
- Molnár, G., Greiff, S., and Csapó, B. (2013). Inductive reasoning, domain specific and complex problem solving: relations and development. *Think. Skills Creat.* 9, 35–45. doi: 10.1016/j.tsc.2013.03.002
- Molnár, G., Greiff, S., Wüstenberg, S., and Fischer, A. (2017). “Empirical study of computer-based assessment of domain-general complex problem-solving skills,” in *The Nature of Problem Solving. Using Research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (Paris: OECD), 125–140.
- Müller, J. C., Kretschmar, A., and Greiff, S. (2013). “Exploring exploration: Inquiries into exploration behavior in complex problem solving assessment,” in *Proceedings of the 6th International Conference on Educational Data Mining*, eds S. K. D’Mello, R. A. Calvo, and A. Olney, 336–337. Available online at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.666.6664&rep=rep1&type=pdf>
- Muthén, L. K., and Muthén, B. O. (2012). *Mplus User’s Guide, 7th Edn*. Los Angeles, CA: Muthén and Muthén.
- Osman, M., and Speekenbrink, M. (2011). Cue utilization and strategy application in stable and unstable dynamic environments. *Cogn. Syst. Res.* 12, 355–364. doi: 10.1016/j.cogsys.2010.12.004
- Schoppek, W., and Fischer, A. (2017). Common process demands of two complex dynamic control tasks: transfer is mediated by comprehensive strategies. *Front. Psychol.* 8:2145. doi: 10.3389/fpsyg.2017.02145
- Siegler, R. S. (1999). Strategic development. *Trends Cogn. Sci.* 3, 430–435.
- Sonnleiter, P., Keller, U., Martin, R., Latour, T., and Brunner, M. (2017). “Assessing complex problem solving in the classroom: meeting challenges and opportunities,” in *The Nature of Problem Solving. Using Research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (Paris: OECD), 159–174.
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learn. Instruct.* 4, 295–312. doi: 10.1016/0959-4752(94)90003-5
- Tein, J. Y., Cox, S., and Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Struct. Equ. Modeling*, 20, 640–657. doi: 10.1080/10705511.2013.824781
- Tóth, K., Rölke, H., Goldhammer, F., and Barkow, I. (2017). “Educational process mining: New possibilities for understanding students’ problem-solving skills,” in *The Nature of Problem Solving. Using Research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (Paris: OECD), 193–210.
- Tschirgi, J. E. (1980). Sensible reasoning: a hypothesis about hypotheses. *Child Dev.* 51, 1–10. doi: 10.2307/1129583
- Vollmeyer, R., Burns, B. D., and Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cogn. Sci.* 20, 75–100. doi: 10.1207/s15516709cog2001\_3
- Wüstenberg, S., Greiff, S., and Funke, J. (2012). Complex problem solving. More than reasoning? *Intelligence* 40, 1–14. doi: 10.1016/j.intell.2011.11.003
- Wüstenberg, S., Stadler, M., Hautamäki, J., and Greiff, S. (2014). The role of strategy knowledge for the application of strategies in complex problem solving tasks. *Technol. Knowledge Learn.* 19, 127–146. doi: 10.1007/s10758-014-9222-8
- Zoanetti, N., and Griffin, P. (2017). “Log-file data as indicators for problem-solving processes,” in *The Nature of Problem Solving. Using Research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (Paris: OECD), 177–191.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Molnár and Csapó. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Potential for Assessing Dynamic Problem-Solving at the Beginning of Higher Education Studies

Benő Csapó<sup>1\*</sup> and Gyöngyvér Molnár<sup>2</sup>

<sup>1</sup> MTA-SZTE Research Group on the Development of Competencies, University of Szeged, Szeged, Hungary, <sup>2</sup> Department of Learning and Instruction, University of Szeged, Szeged, Hungary

## OPEN ACCESS

### Edited by:

Magda Osman,  
Queen Mary University of London,  
United Kingdom

### Reviewed by:

Meredith Ria Wilkinson,  
De Montfort University,  
United Kingdom  
Caroline Di Bernardi Luft,  
Queen Mary University of London,  
United Kingdom

### \*Correspondence:

Benő Csapó  
csapo@edpsy.u-szeged.hu

### Specialty section:

This article was submitted to  
Cognitive Science,  
a section of the journal  
Frontiers in Psychology

**Received:** 02 May 2017

**Accepted:** 06 November 2017

**Published:** 20 November 2017

### Citation:

Csapó B and Molnár G (2017)  
Potential for Assessing Dynamic  
Problem-Solving at the Beginning of  
Higher Education Studies.  
Front. Psychol. 8:2022.  
doi: 10.3389/fpsyg.2017.02022

There is a growing demand for assessment instruments which can be used in higher education, which cover a broader area of competencies than the traditional tests for disciplinary knowledge and domain-specific skills, and which measure students' most important general cognitive capabilities. Around the age of the transition from secondary to tertiary education, such assessments may serve several functions, including selecting the best-prepared candidates for certain fields of study. Dynamic problem-solving (DPS) is a good candidate for such a role, as tasks that assess it involve knowledge acquisition and knowledge utilization as well. The purpose of this study is to validate an online DPS test and to explore its potential for assessing students' DPS skills at the beginning of their higher education studies. Participants in the study were first-year students at a major Hungarian university ( $n = 1468$ ). They took five tests that measured knowledge from their previous studies: Hungarian language and literature, mathematics, history, science and English as a Foreign Language (EFL). A further, sixth test based on the MicroDYN approach, assessed students' DPS skills. A brief questionnaire explored learning strategies and collected data on students' background. The testing took place at the beginning of the first semester in three 2-h sessions. Problem-solving showed relatively strong correlations with mathematics ( $r = 0.492$ ) and science ( $r = 0.401$ ), and moderate correlations with EFL ( $r = 0.227$ ), history ( $r = 0.192$ ), and Hungarian ( $r = 0.125$ ). Weak but still significant correlations were found with certain learning strategies, positive correlations with elaboration strategies, and a negative correlation with memorization strategies. Significant differences were observed between male and female students; men performed significantly better in DPS than women. Results indicated the dominant role of the first phase of solving dynamic problems, as knowledge acquisition correlated more strongly with any other variable than knowledge utilization.

**Keywords:** dynamic problem-solving, technology-based assessment, predictive validity, university admissions, learning strategies

## INTRODUCTION

The social and economic developments of the past decades have re-launched the debate on the mission of schooling, more specifically, on the types of skills schools are expected to develop in their students in order to prepare them for an unknown future. One of the most characteristic features of these debates is a search for a new conception of the knowledge and skills students

are expected to master (see e.g., Adey et al., 2007; Binkley et al., 2012; Greiff et al., 2014). These developments and expectations have reached higher education as well, and novel assessment needs have emerged to reflect the changes. Tests in higher education have traditionally been used as a part of the selection processes (entrance examinations) and to assess students' level of mastery, mostly in the form of summative tests based on the disciplinary content of courses. Recently, the functions of assessments have significantly expanded, thus requiring a renewal of assessment processes in a number of dimensions.

This study lies at the intersection of three rapidly developing fields of research on higher education. The context of the research is set by the practical needs of (1) developing new assessment methods for higher education, including innovative and efficient selection processes for choosing students for higher education studies, and assessing university outcomes beyond disciplinary knowledge and domain-specific skills. These demands have directed the attention of researchers to (2) the twenty-first-century skills as desired outcomes of higher education. Rapidly developing technology-based assessment has made it possible to measure several twenty-first-century skills and to include them in large-scale assessments. (3) Dynamic problem-solving (DPS) is one of those skills which by now has an established research background and may satisfy the needs of higher education. Solving problems in the process of being assessed in DPS based on computer-simulated scenarios involves the component skills of scientific reasoning, knowledge acquisition and knowledge utilization, all necessary for successful higher education studies (see e.g., Buchner and Funke, 1993; Funke, 2001; Greiff et al., 2012; Csapó and Funke, 2017a; Funke and Greiff, 2017). The processes of solving problems in computer-simulated scenarios involve the component skills of scientific reasoning, knowledge acquisition and knowledge utilization, all necessary for learning effectively in higher education (see e.g., Buchner and Funke, 1993; Funke, 2001; Greiff et al., 2012; Csapó and Funke, 2017a; Funke and Greiff, 2017).

As the main construct explored in the present study, DPS has already been defined and assessed in several previous studies. The problems are built on formal models represented by a finite-state automaton, where the output signals are determined by the input signal (Buchner and Funke, 1993). "In contrast to static problems, computer-simulated scenarios provide the unique opportunity to study human problem-solving and decision-making behavior when the task environment changes concurrently to subjects' actions. Subjects can manipulate a specific scenario via a number of input variables [...] and they observe the system's state changes in a number of output variables. In exploring and/or controlling a system, subjects have to continuously acquire and use knowledge about the internal structure of the system" (Blech and Funke, 2005, p. 3).

In the present study, DPS was assessed with a computerized solution based on the MicroDYN approach (Greiff and Funke, 2009; Funke and Greiff, 2017) similar to that employed in delivering most interactive items in the innovative domain in PISA 2012. That assessment framework defined problem-solving in a more general way: "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, 2013a, p. 122). An interpretation of this definition follows: "What distinguishes the 2012 assessment of problem-solving from the 2003 assessment is not so much the definition of problem-solving competency, but the mode of delivery of the 2012 assessment (computer-based) and the inclusion of problems that cannot be solved without the solver interacting with the problem situation" (OECD, 2013a, p. 122). The PISA 2012 problem-solving assessment included both static and interactive tasks, and in this context interactivity is defined as "Interactive: not all information is disclosed; some information has to be uncovered by exploring the problem situation" (OECD, 2014, p. 31, Fig. V.1.2). In the present study, all items are interactive, so the construct we assess is identical with the one PISA assessed in 2012 with its interactive items.

## THEORETICAL FRAMEWORK

### Context of the Study: Need for New Assessments in Higher Education

The need to develop new assessment instruments for higher education has emerged both at international and national levels in a number of countries. There is a general intention to adapt the content of the assessments to changed expectations of the outcomes of higher education. The altered content may then require new assessment methods (see e.g., Bryan and Clegg, 2006). There is a change in the purpose of assessments as well as a visible intention to introduce the principles of evidence-based decision-making and accountability processes to higher education (Hutchings et al., 2015; Ikenberry and Kuh, 2015; Zlatkin-Troitschanskaia et al., 2015). The new functions of assessment go beyond the usual applications of summative tests to measure the mastery level of courses and include estimating educational added value of particular phases of studies, or entire training programs. As there is a great variety of competencies that are outcomes of higher education, thus limiting inter-institutional comparisons in terms of domain-specific competencies, we see a growing need to measure and compare domain-general competencies.

These intentions are clearly marked by feasibility studies launched by the OECD to compare the achievement of college and university students in a number of countries (Assessment of Higher Education Learning Outcomes, AHELO). The AHELO program included assessment of domain-specific competencies as well as of generic cognitive skills, for which the test tasks were adapted from the Collegiate Learning Assessment instrument (Tremblay et al., 2012). Another international initiative, the TUNING CALOHEE project (Measuring and Comparing Achievements of Learning Outcomes in Higher Education in Europe), intends to create an assessment system to compare the outcomes of universities in Europe (Coates, 2016).

In the United States, as the century-long history of successfully administering the Scholastic Aptitude Test (SAT) indicates, admissions processes have always been based on assessing generic

cognitive skills (Atkinson and Geiser, 2009). As studies show, the SAT tests predict achievement in higher education beyond the high school grade point average. They comprise mathematical and verbal components (factor analysis with a recent version of it confirmed the two-factor model, see Wiley et al., 2014), while university admissions in many countries have usually been based on assessing domain-specific competencies (Zlatkin-Troitschanskaia et al., 2015).

A closer context of the study is Hungarian higher education and the admissions process used by its institutions. As there is no specific entrance examination, admissions are based on matriculation examination results. The matriculation examination, like so many other European countries, was introduced in Hungary in the mid-nineteenth century, and it has changed relatively little during its long history. At present, there are three mandatory subjects: (1) Hungarian language and literature, (2) mathematics, and (3) history. Beyond these, students must choose a further subject out of a large number of electives. An examination can be taken at two levels in any subject; there is an intermediate and an advanced exam. There is no exact (measurable) definition for the differences between the two levels. Intermediate exams are taken at students' schools before committees formed from teachers in their own schools, while the advanced exams are centralized and are taken before (independent) committees formed from teachers in other schools. The admissions scores are computed by complex formulas; for advanced exams, extra scores are awarded, and other factors may also be taken into account.

The inadequacy of such a selection criterion is widely discussed, but few research results are available to make evidence-based judgments about the validity of the current practice and about potential alternative solutions. It seems possible that a reformed matriculation examination could serve to certify completion of secondary studies and at the same time could act as a major component of the admissions process (Csapó, 2009). Such a matriculation examination should measure students' knowledge at one level but on a scale which represents a broad range of achievement, should be a technology-based assessment (possibly using item banks and adaptive testing), and should include a few (probably five) compulsory subjects without electives.

The new admissions processes are expected to provide a better prediction of students' success in a changed world of higher education than those of the traditional methods introduced so many decades ago. Assessment of generic cognitive skills, possibly a representative member of the twenty-first-century skills, could also be a component of a new admissions process. To explore the feasibility and validity of such an admissions model, we have measured five domain-specific competencies plus dynamic problem-solving, and we report the results in the present study.

## Definition and Technology-Based Assessment of Twenty-First-Century Skills in Educational Settings

A number of studies have analyzed the requirements of the knowledge-based economy and concluded that science,

technology, engineering, and mathematics (STEM) education should be strengthened and that skills relevant to a dynamically changing technology-rich environment should be developed. In this context, societies today and in the foreseeable future are characterized by a new group of skills, which are often called the twenty-first-century skills, or, in other contexts, transversal skills (Greiff et al., 2014). This loosely defined set of skills includes problem-solving, information and communication skills, critical thinking, creativity, entrepreneurship and collaboration. The topic of twenty-first-century skills has become popular in the literature on the future of education (Trilling and Fadel, 2009; National Research Council, 2013; Kong et al., 2014), and a number of projects have been launched to define, assess, and develop these skills.

Although most skills identified under this label are not new in the sense that they have not been studied before or that they have not been relevant in everyday life, the way they are utilized in this century may be novel. The main novelty is that these skills today are mostly used in a technology-rich environment. Therefore, they should be measured by means of technology. This approach is demonstrated by the Assessment and Teaching of twenty-first-Century Skills (ATC21S) project, among other studies. The first phase of the ATC21S project dealt with definitions and psychometric, technological and policy issues (Griffin et al., 2012), while the second phase focused on the assessment of collaborative problem-solving (Griffin and Care, 2015).

Technology-based assessment has a number of advantages over traditional paper-and-pencil tests in a number of respects. Computerized tests, especially assessments delivered online, may make the entire assessment process more reliable and valid, faster, easier, and less expensive. Beyond these general benefits, there are some constructs which could not be measured without computers. There are domains where technology use is central to the definition of the domain (e.g., information-communication literacy and digital reading), while in other cases it would not be possible to implement the assessment process without technology (Csapó et al., 2012). DPS is such a construct, as students interact with computer-simulated systems during the testing process. Technology is the best means not only to assess these skills, but to develop them as well; for example, simulation- and game-based learning may provide an authentic learning environment to practice these skills (see Qian and Clark, 2016).

Those projects whose aim it was to precisely identify the twenty-first-century skills were able to define only a few of them in a measurable format (Binkley et al., 2012). Even fewer of those skills have an established research background that makes it possible to use them in a large-scale project. Of these, problem-solving, both dynamic (Greiff et al., 2014) and collaborative (OECD, 2013b; Griffin and Care, 2015; Neubert et al., 2015), is sufficiently developed for broader practical use. Beyond these strengths, DPS is a good representative of the twenty-first-century skills because, through its component skills, it may overlap with several other complex skills in this group.

## Assessment of Dynamic Problem-Solving

Problem-solving is one of the most commonly noted constructs among the "new" twenty-first-century skills; it also has a long history in cognitive research (see Fischer et al., 2017). By now,



cognitive research has identified a number of different types of problem-solving which can be classified by several aspects. Domain-specific problem-solving can be distinguished from the domain-general kind, analytical from complex, and static from interactive. In the present study, we deal with the assessment of dynamic problem-solving, which is interactive and can be considered as a specific form of complex problem-solving. Dynamic problem-solving, as was shown in the previous section, can only be measured by means of technology.

Complex problem-solving has already been studied in a number of contexts; previous research shows that it is a generic cognitive skill, but is different from general intelligence (Funke, 2010; Wüstenberg et al., 2012; Greiff et al., 2013a). Using computers to assess problem-solving has allowed a migration of previous paper-based tests to an electronic platform, thus improving the efficiency and usability of the tests as well as opening up a range of new prospects (Wirth and Klieme, 2003). These new possibilities include constructing more real life-like scenarios, using simulations, offering interactive activities, and in this way improving the ecological validity of the assessments in general.

Using simulation to study problem-solving was already proposed long ago (Funke, 1988), but the broad availability of computers launched a new wave of research based on computer-simulated systems (Funke, 1993, 1998; Greiff et al., 2013a). The difficulty level of tasks based on simulation is easily scalable; even simulated minimal complex systems offer outstanding opportunities to study the processes of problem-solving (Sonnleitner et al., 2012; Funke, 2014; Funke and Greiff, 2017; Greiff and Funke, 2017).

DPS as a specific category of complex, interactive problem-solving offers outstanding potential both to create tests for laboratory studies and for large-scale assessment (Buchner and Funke, 1993; Funke, 2001; Greiff et al., 2012). When students solve dynamic problems on a computer, their activities can be logged and fine mechanisms of their cognition can be explored by analyzing log files (Tóth et al., 2017).

Several types of problem-solving have already been assessed three times within the framework of PISA. First, static problem-solving was assessed in 2003 with paper-based tests (OECD, 2004; Fleischer et al., 2017). Then, in 2012, problem-solving was measured with computerized tests comprising two types of tasks, static (15 items) and interactive (27 items). The static items were similar to those of the PISA 2003 assessment; they were computerized versions of items that would be possible to measure with paper-and-pencil tests as well, while the interactive items were novel in large-scale assessments and measured the same construct (DPS) as the present study, based on the MicroDYN approach, too (Greiff and Funke, 2009; Funke and Greiff, 2017).

The 2012 PISA assessment was the first large-scale assessment of DPS in international context and demonstrated that there were large differences between the participating countries in the problem-solving performance of their students, even if the achievement on the main literacy domains was similar (OECD, 2014). The successful completion of the 2012 PISA DPS assessment has accelerated research in this field and inspired a number of further studies (see Csapó and Funke,

2017a). In PISA 2015, collaborative problem-solving was the innovative domain; collaboration was simulated by human-agent interactions (OECD, 2013b).

Assessments of problem-solving have already proved useful in higher education, but the vast majority of them covered domain-specific problem-solving (e.g., Lopez et al., 2014; Zlatkin-Troitschanskaia et al., 2015). As technology-based assessment instruments become more widely available, such skills have been measured more often, capitalizing on experiences from the computer-based assessment of problem-solving. These assessments may be especially useful when the cognitive outcomes of some innovative instructional methods are measured, such as project methods, problem-based learning and inquiry-based learning.

In the present study, we go further when we explore the possibilities for assessing domain-general problem-solving. Our test is based on the MicroDYN approach (Greiff and Funke, 2009; Funke and Greiff, 2017), which measures the same construct that was measured with several dynamic items in the PISA 2012 assessment (OECD, 2014) and in several other studies (Abele et al., 2012; Molnár et al., 2013, 2017; Frischkorn et al., 2014).

DPS tasks have the same general characteristics. Simulated systems are presented which are based on practical contexts and situations that are easy for the problem-solver to comprehend. The simulated systems show a well-defined behavior, the problem-solver has to manipulate some input variables, and the system responds with changes in output variables. This represents a major difference over paper-based tests, as this sort of a realistic interaction with a responding system cannot be created on paper. The purpose of the interaction is to comprehend the rules that determine the behavior of the system.

In the first phase of completing a DPS test task, students interact with the simulated system, manipulate the values of independent variables, and observe how the changes impact the values of dependent variables. This interactive observation is the *knowledge acquisition phase* (also referred to as the rule identification phase), after which students depict the results of their observations on a concept map. Then, they have to manipulate the variables so that they reach a goal state; this is the knowledge utilization phase (or rule application phase). The results from the two phases are scored separately, and as previous research (e.g., Wüstenberg et al., 2012) has shown, there may be significant differences in performance in the two phases. These dynamic tasks are easily scalable, as the number of input and output variables as well as the relationships between them can be changed.

## Aims and Research Questions

In the present study, we explore the prospects and value of assessing DPS in higher education. Such a test could later be a useful component of university admissions processes, especially in STEM disciplines, where studies require problem-solving in a technology-rich environment. The context of the study allows for an examination of the relationships between subject matter knowledge and problem-solving.

- RQ1: How do matriculation examination results predict problem-solving test performance assessed at the beginning of higher education studies? We assume that the knowledge students possessed at the end of their high school studies (assessed by the matriculation exams) correlates with problem-solving, but the strengths of the relationships with the different domains is still open.
- RQ2: What are the relationships between subject matter tests and problem-solving performance measured at the beginning of higher education studies? We assume that although the disciplinary knowledge tests correlate well with problem-solving, problem-solving measures other aspects of knowledge; therefore, it has considerable added value over subject matter tests.
- RQ3: Are there differences between students in different disciplines? We assume that those who study within different divisions at the university also differ in their problem-solving skills and in the relationships between problem-solving and other tests.
- RQ4: How do students' characteristics and background variables influence their problem-solving performance? We may assume that students' family background, mother's level of education, students' intention to learn and students' learning strategies influence how their problem-solving skills develop.

We expect that the results from these analyses may contribute to improving matriculation examinations as well as to devising better admissions processes.

## METHODS

### Participants

Participants in the study were students admitted to a large Hungarian university and starting their studies. The university has 12 divisions (arts, science, medicine, etc.), but they vary in size (number of students). All of the divisions participated in the study, but because of the differences between them, not all analyses are equally relevant for every division.

The population for the study was formed exclusively of students who had just finished their high school studies and immediately applied for admission to the university. They took their matriculation examinations in May, and the assessment for this study was carried out in September of the same year.

The target population was 2,319 students, of whom 1,468 (63.3%) participated in the assessment; 57.7% of them were female. The participation rate by division varied from 28.18 to 74.16%.

Student participation was voluntary; students were notified of their option to take part in the assessment prior to commencing their studies. As an incentive, they received credits for successful completion of the tests.

### Instruments

#### Problem-Solving Test

Students completed a DPS test based on the MicroDYN approach. Several tests composed of similar tasks based on this

model have already been used in other studies in Hungary (Molnár et al., 2013, 2017) but only with younger participants. The test prepared for this study consisted of 20 items with varying difficulty levels.

For example, in the knowledge acquisition phase of an easy item, students had to observe how changing the values of two independent variables (e.g., two different kinds of syrup) impacted the value of one dependent (target) variable (sweetness of the lemonade). They moved sliders on the screen to set the current value for the blue and for the green syrup. The system responded by indicating the resultant sweetness level. Students observed what happened and attempted a new setting, observing the sweetness level with such a setting. They had 180 s for the knowledge acquisition phase in each task. In the knowledge utilization phase, they had to reach the required value of the dependent variable (sweetness) by setting the proper values of the independent variables in no more than 180 s. In a difficult item, students had to comprehend more complex relationships between three independent variables (three different training methods used by basketball players) and three dependent variables (motivation, power of the throw and exhaustion). (For more examples of similar DPS items, see Greiff et al., 2013b; OECD, 2014). The two phases of problem-solving were scored separately. The score for the first (knowledge acquisition) phase was based on how accurately the relationships between the variables were depicted, while the score for the second (knowledge utilization) phase reflected the success with which the dependent variables reached the target state.

The difficulty level of the test was close to the optimal for the whole sample with a 45% mean ( $SD = 21.74$ ). The reliability (Cronbach's alpha) of the entire test was 0.88. The reliabilities of the two problem-solving phases were also high (knowledge acquisition: 0.84; knowledge utilization: 0.83).

#### Disciplinary Knowledge Tests

Five disciplinary knowledge tests were prepared for the assessment: Hungarian language and literature (Hungarian, for short, with a strong reading comprehension component), mathematics, history, science, and English as a Foreign Language (EFL). Test content was based on the students' high school studies. The tests covered the major topics of the particular disciplines. Difficulty levels for the tests were adjusted approximately to the intermediate-level standards of the matriculation examination. These tests were prepared by experts practiced in preparing matriculation examination tests. The tests made use of the options made available by computer-based testing; using a variety of stimuli (e.g., texts, images, and animation) and response capture (e.g., entering texts and numbers, clicking, and moving objects on the screen by drag-and-drop). The descriptive statistics for the entire sample and the reliability of the instruments are summarized in **Table 1**. The reliability coefficients for the tests were good, ranging from 0.88 to 0.96.

#### Background Questionnaire

A background questionnaire was administered to participating students via the same platform as the tests. Data were collected

in this way about their matriculation examination results and their learning strategies and SES. To minimize the time devoted to administering the questionnaire, only the most relevant variables were explored, where strong relationships were expected. Family background was represented by mothers' level of education (from primary school to master's degree). Students' commitment to study (intention to learn) was measured with the highest degree they intend to earn (bachelor's, master's or PhD). Two scales for learning strategies that use self-reported Likert scales were adapted from the PISA 2000 assessment (elaboration strategies and memorization strategies, see Artelt et al., 2003).

## Procedures

The assessments were carried out in a large computer room at the university learning and information center. Three 2-h sessions (1h per test) were offered to the students in the first 2 weeks of the semester. The tests and the questionnaire were administered using the eDia online platform.

Students received detailed feedback on their performance a week after the testing period ended. The feedback contained detailed analyses of their performance in the context of normative comparative data.

Data from the achievement tests were analyzed with IRT models. Plausible values were computed to compare the achievements of the age groups, and Weighted Likelihood Estimates (WLE) were used to compute person parameters. The analyses were performed with the ACER ConQuest program package (Wu et al., 1998). Person parameters were transformed to a 500(100) scale so that the university means were set to 500. MPlus software was used to conduct the structural equation modeling (Muthén and Muthén, 2012).

## RESULTS

In this section, we first answer the research questions by examining the details of the correlations between subject matter knowledge represented in the matriculation examination results and in the test scores from the beginning of studies in higher education. Then, we synthesize the relationships in a path model based on these findings.

**TABLE 1 |** Disciplinary knowledge test: descriptive statistics and reliability coefficients.

Test	Number of items	Mean (%)	Standard deviation	Cronbach's alpha
Hungarian	126	34.30	9.38	0.90
Mathematics	63	59.79	14.27	0.89
History	161	58.60	11.91	0.93
Science	163	45.31	9.01	0.88
EFL	80	55.70	19.11	0.96

## Matriculation Examination Results as Predictors of Problem-Solving Performance

Performance in two phases of problem-solving (knowledge acquisition and knowledge utilization) correlated at the moderate level ( $r = 0.432$ ,  $p < 0.001$ ); therefore, it is worth examining the correlations between the matriculation examination results and the phases of problem-solving separately. Here, we only deal systematically with the three mandatory matriculation examination subjects, as these data are available for all participants, while only a small proportion of students took the exams in a science discipline or EFL as an elective. As few students took the matriculation examinations at the advanced level, this analysis involves the results from the intermediate exams. For a comparison, we have computed the correlations between matriculation examination results and those from the knowledge tests (see Table 2).

Two major observations stand out from Table 2. First, the mathematics matriculation result (which is based on a paper-and-pencil test with constructed responses) predicts problem-solving much more strongly than those in the two other subjects. Second, knowledge acquisition has a stronger correlation with the matriculation examination results than knowledge utilization does. The mathematics and history matriculation results predict the test results for the same respective subjects well; they are lower for Hungarian, which has no significant correlation with problem-solving. We note that when comparing the correlations, ca.0.05 differences are significant at  $p < 0.05$ , while ca.0.1 differences are significant at  $p < 0.001$  (one-tailed, calculated by the Fisher  $r$ -to- $z$  transformation). When we note differences between correlations, they are statistically significant.

## Relationships between Subject Matter Tests and Problem-Solving

Correlations for the six tests are summarized in Table 3. The correlations between disciplinary knowledge test results are moderate (Hungarian and history with science and EFL) or large, and as expected from the similarities between these subjects, the Hungarian–history and mathematics–science pairs correlate more strongly than other pairs. Mathematics has the strongest correlation with problem-solving, followed by science.

These correlations confirm once again that knowledge acquisition is a more decisive component of problem-solving items than knowledge utilization. To examine the details of this

**TABLE 2 |** Correlations between the matriculation examination results and those from the tests administered at the beginning of higher education studies.

Matriculation examination	Tests					
	Hungarian	Mathematics	History	Knowledge acquisition	Knowledge utilization	Problem-solving
Hungarian	0.378***	0.071*	0.220***	n.s.	n.s.	n.s.
Mathematics	0.291***	0.656***	0.233***	0.426***	0.273**	0.414***
History	0.395***	0.219***	0.503***	0.133***	n.s.	0.109**

\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ .

relationship, we performed regression analyses with problem-solving and its two phases as dependent variables, using the disciplinary knowledge test results as independent variables (Table 4).

The differences between these analyses confirm previous observations on the role of knowledge acquisition and indicate that it is only mathematics and science whose contribution to the variance explained is significant and positive. Furthermore, even in the cases of knowledge acquisition, ~70% of the variance remained unexplained.

## Differences between Students Studying within Different Divisions

As can be expected, there are large differences between the divisions at the university, both in performance on knowledge tests and on problem-solving. Therefore, it is anticipated that problem-solving has different relationships with disciplinary knowledge. To examine these differences, we have chosen two divisions with a large number of students participating in the assessments and with different study profiles. The division that deals with the humanities, known as the Faculty of Arts (Arts, for short), participated with 212 students (65.2% of the population, 71.7% female), and the division that deals mainly with the natural sciences, known as the Faculty of Science and Informatics (Science, for short), was represented with 380 students (64.0%

of the population, 32.8% female). They performed differently on each test (Table 5), including problem-solving.

Achievement differed according to the expectations for the different study profiles. Students at the Arts Faculty performed better in Hungarian, history and EFL, while Science Faculty students performed better in mathematics, science and problem-solving.

To examine the details of the relations between disciplinary knowledge and problem-solving, we performed the regression analyses separately for the two divisions. Taking into account the decisive role of knowledge acquisition, we present only the results for this phase of problem-solving in Table 6. For comparison, the  $R^2$  were 0.203 (Arts) and 0.217 (Science) for the entire problem-solving test when the same analyses were performed.

Although the same amount of variance of knowledge acquisition was explained by the same set of independent variables, the contributions of the individual variables are different. Mathematics and science play an important role at both divisions, and the contribution of EFL is also significant at the Faculty of Science.

## Relationships between Students' Background Variables and Problem-Solving Performance

Previous studies (e.g., OECD, 2014) have indicated large difference in problem-solving in a number of dimensions. Here, we explore the differences according to some available background variables.

**TABLE 3 |** Correlations for the tests taken at the beginning of higher education studies.

Subject	Hungari-an	Mathematics	History	Science	EFL
Mathematics	0.434***				
History	0.598***	0.409***			
Science	0.375***	0.529***	0.395***		
EFL	0.307***	0.341***	0.337***	0.399***	
Knowledge acquisition	0.156**	0.515***	0.228***	0.422***	0.262***
Knowledge utilization	n.s.	0.315***	0.095**	0.254***	0.121**
Problem-solving	0.125**	0.492***	0.192***	0.401***	0.227***

\*\* $p < 0.01$ , \*\*\* $p < 0.001$ .

**TABLE 5 |** Differences in achievement among students in the two divisions with different study profiles.

Division	Hunga- rian	Mathe- matics	History	Science	EFL	Problem- solving
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Arts	539 (98)	466 (92)	529 (99)	481 (78)	526 (108)	464 (94)
Science	486 (89)	546 (103)	495 (94)	525 (91)	506 (94)	542 (93)
<i>t</i>	6.28	−9.59	3.87	−5.87	2.02	−9.38
Sig	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.05$	$p < 0.001$

**TABLE 4 |** Regression analyses of problem-solving and its two phases as dependent variables with disciplinary knowledge tests as independent variables.

Independent variables	Problem-solving			Knowledge acquisition			Knowledge utilization		
	$R^2 = 0.286$			$R^2 = 0.309$			$R^2 = 0.121$		
	Beta	<i>t</i>	Sig	Beta	<i>t</i>	Sig	Beta	<i>t</i>	Sig
Hungarian	−0.151	−4.896	0.000	−0.141	−4.628	0.000	−0.115	−3.351	0.001
Mathematics	0.420	14.056	0.000	0.423	14.424	0.000	0.284	8.585	0.000
History	0.013	0.429	0.668	0.035	1.145	0.252	−0.013	−0.380	0.704
Science	0.218	7.257	0.000	0.214	7.238	0.000	0.154	4.622	0.000
EFL	0.036	1.329	0.184	0.060	2.265	0.024	0.000	−0.010	0.992



**TABLE 6 |** Regression analyses of knowledge acquisition as a dependent variable with disciplinary knowledge tests as independent variables for the two divisions.

Independent variables	Faculty of Arts			Faculty of Science		
	$R^2 = 0.237$ ( $p < 0.001$ )			$R^2 = 0.237$ ( $p < 0.001$ )		
	Beta	t	Sig	Beta	t	Sig
Hungarian	0.109	1.343	0.181	-0.143	-2.242	0.026
Mathematics	0.252	3.057	0.003	0.379	6.444	0.000
History	0.065	0.806	0.421	-0.035	-0.577	0.564
Science	0.187	2.340	0.020	0.154	2.789	0.006
EFL	0.006	0.087	0.931	0.165	3.072	0.002

### Gender Differences

Gender differences are routinely analyzed on large-scale national and international assessments. The PISA studies indicated that Hungarian girls' reading comprehension was significantly better than that of boys, while boys' performance was better in mathematics and there were no significant gender differences in science (OECD, 2016). Female and male students performed differently on problem-solving in this study as well. To provide context to interpret the size of gender difference in problem-solving, the differences on other tests are also indicated in Table 7.

The only test where women outperform men was Hungarian language (in line with the better reading performance of the female students); on all other tests, men performed better. The largest difference was found in favor of men in problem-solving. Here again, knowledge acquisition shows a much larger difference, indicating that this is the more sensitive phase of problem-solving.

### Mothers' Education and Intention to Learn

The relationship of test performance with students' socio-economic status is a well-known phenomenon, although there are large differences in this respect between countries and also between domains of assessment. International assessment programs (e.g., the PISA studies) usually involve complex indices for this purpose, but we have only one variable to represent students' family background, mothers' education. A further variable that may be interesting in this context is what degree students want to earn (intention to learn). We have found a small (Spearman's  $\rho = 0.182$ ,  $p < 0.001$ ) correlation between these two variables. The correlations of test results with mother's educational level and intention to learn are summarized in Table 8.

There are no large differences between the correlations; all are rather small. Mothers' education has little impact on problem-solving. The correlation of problem-solving with intention to learn is small but still significant; the correlation with knowledge utilization here is also smaller than with knowledge acquisition.

### Learning Strategies

As there are only a few questions in the learning strategies questionnaire, we present the texts and the correlations with

**TABLE 7 |** Gender differences in test performance.

Test	Gender	Mean	SD	Difference	t
Hungarian	M	489	106	-26	-4.54
	F	515	94		
Mathematics	M	534	102	49	8.75
	F	485	94		
History	M	524	105	39	7.03
	F	485	92		
Science	M	518	95	29	5.18
	F	489	101		
EFL	M	513	102	23	3.93
	F	490	97		
Knowledge acquisition	M	597	108	93	14.86
	F	503	111		
Knowledge utilization	M	492	126	62	9.40
	F	430	100		
Problem-solving	M	545	98	78	14.6
	F	467	88		

All differences are significant at  $p < 0.01$ .

**TABLE 8 |** Correlations of performance on the tests with mother's education and intention to learn.

Test	Mother's education	Intention to learn
Hungarian	0.153***	0.189***
Mathematics	0.145***	0.167***
History	0.134***	0.190***
Science	0.161***	0.242***
EFL	0.192***	0.157***
Knowledge acquisition	0.084**	0.116***
Knowledge utilization	n.s.	0.064**
Problem-solving	n.s.	0.105***

\*\* $p < 0.01$ , \*\*\* $p < 0.001$ .

the problem-solving achievement for each question. Students' answers to these questions show small but significant correlations with problem-solving (Table 9).

The elaboration strategies questions correlate positively with problem-solving, while the memorization strategy questions correlate negatively with it. It is quite clear from the content of the questions that students who prefer conceptual meaningful learning over rote learning are better problem-solvers.

## An Integrated Model of the Relations of Knowledge Acquisition in Dynamic Problem-Solving

We synthesized the results using structural equation modeling (SEM). Taking into account the observations reported in the previous sections, here we deal only with the knowledge acquisition phase. As the main aim of the present study is to validate the DPS test and to explore its usefulness at the beginning of university studies, we conceived a model by using

**TABLE 9 |** Correlations of the learning strategies questions with problem-solving performance.

Question	DPS1	DPS2	DPS
<b>ELABORATION STRATEGIES</b>			
When I study, I try to relate new material to things I have learned in other subjects.	0.129***	0.075**	0.121***
When I study, I figure out how the information might be useful in the real world.	n.s.	n.s.	n.s.
When I study, I try to understand the material better by relating it to things I already know.	0.080**	0.104***	0.109***
When I study, I figure out how the material fits in with what I have learned.	n.s.	0.070*	0.069*
<b>MEMORIZATION STRATEGIES</b>			
When I study, I try to memorize everything that might be covered.	-0.153***	-0.091**	-0.144**
When I study, I memorize as much as possible.	-0.097**	n.s.	-0.074**
When I study, I memorize all new material so that I can recite it.	-0.183***	-0.128***	-0.185**
When I study, I practice by saying the material to myself over and over.	-0.263***	-0.135***	-0.236**

DPS1, Knowledge acquisition; DPS2, Knowledge utilization; DPS, dynamic problem-solving. \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$ .

variables with significant correlations. We assume that students' gender and learning strategies influence their disciplinary test results, while these results (students' actual knowledge) influence achievement in DPS.

A model that adequately fits the data (RMSEA = 0.046, CFI = 0.986, TLI = 0.949) is presented in **Figure 1**. Gender influences mathematics test results, while learning strategies have a remarkable impact on mathematics and science. These two disciplines and history have a significant relationship with the first phase of problem-solving.

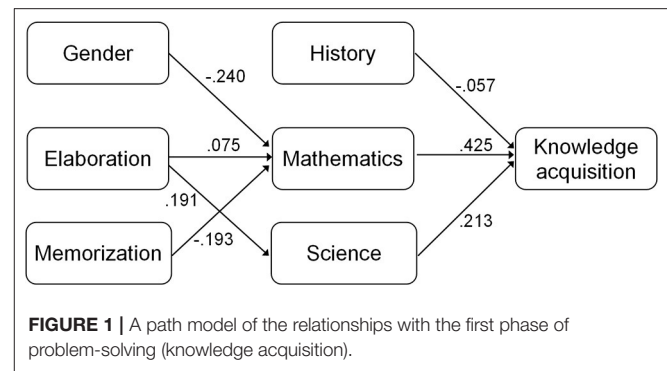
In this model, positive impacts of elaboration strategies on mathematics and science were found, while success in the knowledge acquisition phase of DPS was positively influenced by science and mathematics. Gender and memorization strategy as well as history have a negative relationship.

## DISCUSSION

Our results confirmed or extended several findings from previous research (e.g., different relationships of the phases of problem-solving) and identified some new relationships as well (e.g., the relationships with learning strategies).

### Determinants of Problem-Solving Achievement at the Beginning of Higher Education Studies

Previous research has already identified several characteristics of DPS at different ages, including primary and secondary school students (Molnár et al., 2013), 15-year-old students in the



PISA 2012 assessment (which used tests partially built on the MicroDYN approach in a large-scale international comparative survey, OECD, 2014), and university students (Wüstenberg et al., 2012). The present study has shown the feasibility and usefulness of such an assessment in higher education, indicating that DPS is an easily applicable test with several characteristics of the twenty-first-century skills.

Based on the available data, the impact of previous learning was represented by the disciplinary knowledge test results of three matriculation examination subjects. Mathematics had the strongest correlation with problem-solving, which can be explained by the fact that mathematics is studied throughout the 12 years of primary and secondary schooling and by the nature of cognitive processes required by problem-solving (Greiff et al., 2013b; Csapó and Funke, 2017b). The important role of mathematics was also noticed when the correlations with the subject tests were analyzed, and the integrating path model mirrored the same exceptional impact as well.

The first phase of solving dynamic items (knowledge acquisition or rule identification in other studies) has a stronger relationship with any other observed variable than the second phase (knowledge utilization or rule application). Other studies have found similar differences, although the dominance of knowledge acquisition was not so obvious (Wüstenberg et al., 2012). The important role of the first phase, indicated by larger correlations, may be attributed to the kind of reasoning this phase requires. Students have to combine the different values of the independent variables they manipulate in this phase (combinatorial reasoning), judge certain probabilities (probabilistic reasoning) and abstract rules from the observed behavior of the simulated system (inductive reasoning). This may also explain the strong connection (especially of the first phase) to mathematics, as this kind of reasoning is mostly applied when learning mathematics. Rule induction connects DPS to general intelligence as well, as most intelligence tests use inductive reasoning items. Nevertheless, previous research has indicated that problem-solving explains added variance of students' school achievement (GPA) beyond intelligence tests (Wüstenberg et al., 2012), and moderate to large correlation ( $r = 0.44, 0.52$ , and  $0.47$  in Grades 5, 7, and 11) has been found between problem-solving and inductive reasoning (Molnár et al., 2013).

Our analyses showed that there were differences between the students preparing for studies in different disciplines both on

the level of problem-solving achievement and in the strengths of correlations with domain-specific knowledge. However, some main tendencies, e.g., the dominant role of mathematics and science and the role of the knowledge acquisition phase, may be generalized.

Large gender differences were found on all the tests we used in this study, but the largest one was observed in problem-solving (78 points), mathematics being the second largest one with a much lower difference (49 points). The difference in knowledge acquisition is especially high (93 points). In PISA 2012, gender differences in problem-solving varied from country to country. The OECD mean was 7 points, and in Hungary it was below average, though it was not significant, only 3 points (OECD, 2014). To interpret this discrepancy between the PISA results and the present study, it is worth noting that the Hungarian PISA problem-solving results were below average (459 points) and that not all items were dynamic. Furthermore, in our study, women are overrepresented in the Arts Faculty and in humanities studies in general, while they are underrepresented in STEM studies.

Although there are large differences between students according to the socio-economic status (SES) of their family—and Hungary belongs to a group of countries where the impact of SES is especially strong—there was no large effect found in problem-solving in the PISA 2012 survey. In our study, we also found a modest impact of mothers' level of education on students' problem-solving performance. The fact that problem-solving is less determined by social background than domain-specific competencies indicates a potential opportunity for disadvantaged students as they may show their strengths on these kinds of assessments.

Previous studies have indicated a strong relationship between low- and high-achieving high school students and the different learning strategies they use (Yip, 2013). Our results confirm this notion, as there are clear links between learning strategies and knowledge acquisition in problem-solving. A positive effect of elaboration strategies may have been predictable, but a measurable negative impact of memorization strategies is somewhat unexpected. These results suggest the conclusion that problem-solving is learnable and point to one of the directions in the search for proper training methods. In general, there are two main directions for facilitating the development of this kind of general cognitive skill in a school context. The first is a holistic approach, when developmental impacts are embedded in other educational activities, in this case in learning science and mathematics through meaningful elaborative strategies. Discovery learning and inquiry-based teaching methods may have an impact on the development of problem-solving as well. The second method improves problem-solving by developing component skills (Csapó and Funke, 2017b). We have identified potential component skills; providing training in them may also influence the development of problem-solving.

The results of the SEM indicate the complex nature of the relationships between the variables being explored. The DPS tasks are constructed so that completing them requires no preliminary knowledge within any discipline. Therefore, we may assume that if there are relationships between disciplinary knowledge tests and DPS tests, these relationships are established

by factors other than the factual knowledge represented in the knowledge tests. Such factors may be learning strategies (we have variables for representing them) and certain cognitive skills needed both for completing the disciplinary tests and the DPS tests (in this study, we have no variables to represent them in the SEM). In this model, gender as a variable (most probably) mediates women's better reading and poorer mathematics achievement (shown by other studies, e.g., PISA). In sum, this model indicates that men outperform women, and this impact is mediated by the higher mathematics performance among men. The negative impact of memorization is transmitted via mathematics and science.

## Limitations of the Present Study

As the PISA 2012 assessments also indicated, there are large differences between countries not only on the level of problem-solving performance, but also in the strengths of the relationships between several relevant variables as well; therefore, some particular results found in one country cannot be generalized over countries and cultures. Although some general tendencies were found, we have seen that the strength of the relationships we have examined in this study differs by division. Therefore, the generalizability of the strengths of these relationships is limited; nevertheless, the method we applied in this study is generalizable and may be useful to explore the actual relationships in any higher educational context. Participation was voluntary in the study; the actual samples are thus not representative of the divisions. Nevertheless, the analyses revealed some generalizable tendencies as well.

## Conclusions: Further Research and Prospects for Assessment Supporting High School–University Transition

The results from the present study have raised several further questions worth researching. The dominant role of knowledge acquisition indicates a promising line of inquiry to explore this phase in more detail. One promising direction is to identify students' knowledge acquisition strategies, e.g., the way they manipulate the independent variables when they attempt to discover how these manipulations impact changes in the dependent variable. Students' activities are logged, and their strategies may be ascertained with log file analyses. Latent class analysis may be an effective method to identify students' exploring strategies.

The knowledge acquisition phase also deserves further study from the perspective of its relationships to learning strategies as well, for example, examining if poor problem-solving performance can be an indicator of inadequate learning strategies. If such a connection can be proven, problem-solving assessment could be a diagnostic tool for identifying poor learning strategies, possibly more reliable than self-reported questionnaires. Further insights into the nature of cognition in the knowledge acquisition phase could be expected from studying it in relation to the learning to learn assessments (Greiff et al., 2013b; Vainikainen et al., 2015).

Several skills may be identified which are needed to successfully complete phases of problem-solving. A systematic examination of the role of some supposed component skills (e.g., combinatorial reasoning, probabilistic reasoning, correlational reasoning and inductive reasoning) would provide foundations for the development of problem-solving by strengthening its component skills as well.

The results from this study indicated that technology-based assessment of problem-solving may be a useful instrument to moderate the secondary–tertiary education transition. To improve its usefulness, the scoring system may be further developed, extending it with an automated log file analysis. Such an instrument would be especially helpful in selection processes (admissions tests) for the STEM disciplines. More detailed analyses of the relationship between problem-solving and the study profile would be needed to improve the test. In the present study, we compared divisions of study within the university, but a division is still not homogeneous; for example, students in biology training may be different from those in mathematics.

We have found significant positive relationships with the questions on elaboration learning strategies and negative relationships with the questions on memorization strategies. In the present study, there were not enough questions to use sophisticated scales for representing these learning strategies,

but the findings indicate the relevance of exploring the role learning strategies play in the development of problem-solving. This seems a promising area both for research and practice not only for higher education but also for earlier phases of school education.

The predictive power of DPS can be explored later when data is available on the university achievement of the students participating in the present assessment. The test may have a diagnostic value (indicating poor study strategies or insufficient problem-solving skills) and can also be used to aid students in selecting a study track better suited to their cognitive skills.

## AUTHOR CONTRIBUTIONS

BC and GM have made an equal contribution to the study, including the design, data collection, analyses, and writing of the manuscript, and have both approved it for publication.

## FUNDING

This study was completed within the research program of the MTA–SZTE Research Group on the Development of Competencies. The data analysis and preparation of the manuscript were funded by OTKA K115497.

## REFERENCES

- Abele, S., Greiff, S., Gschwendtner, T., Wüstenberg, S., Nickolaus, R., Nitzschke, A., et al. (2012). Dynamische Problemlösekompetenz. Ein bedeutsamer Prädiktor von Problemlöseleistungen in technischen Anforderungskontexten? [Dynamic problem solving – an important predictor of problem-solving performance in technical domains?]. *Zeitschrift für Erziehungswissenschaft* 15, 363–391. doi: 10.1007/s11618-012-0277-9
- Adey, P., Csapó, B., Demetriou, A., Hautamaki, J., and Shayer, M. (2007). Can we be intelligent about intelligence? Why education needs the concept of plastic general ability. *Educ. Res. Rev.* 2, 75–97. doi: 10.1016/j.edurev.2007.05.001
- Artelt, C., Baumert, J., Julius-McElvany, N., and Peschar, J. (2003). *Learners for Life. Student Approaches to Learning. Results from PISA 2000*. Paris: OECD.
- Atkinson, R. C., and Geiser, S. (2009). Reflections on a century of college admissions tests. *Educ. Res.* 38, 665–676. doi: 10.3102/0013189X09351981
- Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., et al. (2012). “Defining twenty-first century skills,” in *Assessment and Teaching of 21st Century Skills*, eds P. Griffin, B. McGaw, and E. Care (New York, NY: Springer), 17–66.
- Blech, C., and Funke, J. (2005). *Dynamis review: An overview about Applications of the Dynamis Approach in Cognitive Psychology*. Bonn: German Institute for Adult Education (DIE). Available online at: [http://www.die-bonn.de/esprid/dokumente/doc-2005/blech05\\_01.pdf](http://www.die-bonn.de/esprid/dokumente/doc-2005/blech05_01.pdf)
- Bryan, C., and Clegg, K. (eds.). (2006). *Innovative Assessment in Higher Education*. Abingdon: Routledge.
- Buchner, A., and Funke, J. (1993). Finite-state automata: dynamic task environments in problem-solving research. *Q. J. Exp. Psychol.* 46, 83–118. doi: 10.1080/14640749308401068
- Coates, H. (2016). Assessing student learning outcomes internationally: insights and frontiers. *Asses. Eval. High. Educ.* 41, 662–676. doi: 10.1080/02602938.2016.1160273
- Csapó, B. (2009). “The second stage of public education and the Matura,” in *Green Book for the Renewal of Public Education in Hungary*, eds K. Fazekas, J. Köllö, and J. Varga (Budapest: Ecostat), 81–104.
- Csapó, B., Ainley, J., Bennett, R., Latour, T., and Law, N. (2012). “Technological issues of computer-based assessment of 21st century skills,” in *Assessment and Teaching of 21st Century Skills*, eds P. Griffin, B. McGaw, and E. Care (New York, NY: Springer), 143–230.
- Csapó, B., and Funke, J. (eds.) (2017a). *The Nature of Problem Solving. Using Research to Inspire 21st Century Learning*. Paris: OECD.
- Csapó, B., and Funke, J. (2017b). “The development and assessment of problem solving in 21st-century schools,” in *The Nature of Problem Solving. Using Research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (OECD: Paris), 19–32.
- Fischer, A., Greiff, S., and Funke, J. (2017). “The history of complex problem solving,” in *The Nature of Problem Solving. Using Research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (OECD: Paris), 107–121.
- Fleischer, J., Buchwald, F., Leutner, D., Wirth, J., and Rumann, S. (2017). “Analytical problem solving: Potentials and manifestations,” in *The Nature of Problem Solving. Using Research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (OECD: Paris), 33–45.
- Frischkorn, G. T., Greiff, S., and Wüstenberg, S. (2014). The development of complex problem solving in adolescence: a latent growth curve analysis. *J. Educ. Psychol.* 106, 1007–1020. doi: 10.1037/a0037114
- Funke, J. (1988). Using simulation to study complex problem solving: a review of studies in the FRG. *Simul. Games* 19, 277–303. doi: 10.1177/0037550088193003
- Funke, J. (1993). Microworlds based on linear equation systems: a new approach to complex problem solving and experimental results. *Adv. Psychol.* 101, 313–330. doi: 10.1016/S0166-4115(08)62663-1
- Funke, J. (1998). Computer-based testing and training with scenarios from complex problem-solving research: advantages and disadvantages. *Int. J. Sel. Asses.* 6, 90–96. doi: 10.1111/1468-2389.00077
- Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Think. Reason.* 7, 69–89. doi: 10.1080/13546780042000046
- Funke, J. (2010). Complex problem solving: a case for complex cognition? *Cogn. Process.* 11, 133–142. doi: 10.1007/s10339-009-0345-0
- Funke, J. (2014). Analysis of minimal complex systems and complex problem solving require different forms of causal cognition. *Front. Psychol.* 5:739. doi: 10.3389/fpsyg.2014.00739



- Funke, J., and Greiff, S. (2017). "Dynamic problem solving: Multiple-item testing based on minimally complex systems," in *Competence Assessment in Education*, eds D. Leutner, J. Fleischer, J. Grünkorn, and E. Klieme (Cham: Springer International Publishing), 427–443.
- Greiff, S., and Funke, J. (2009). "Measuring complex problem solving: The MicroDYN approach," in *The Transition to Computer-Based Assessment. New Approaches to Skills Assessment and Implications for Large-Scale Testing*, eds F. Scheuermann, and J. Björnsson (Luxembourg: Office for Official Publications of the European Communities), 157–163.
- Greiff, S., and Funke, J. (2017). "Interactive problem solving: Exploring the potential of minimal complex systems," in *The Nature of Problem Solving. Using Research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (Paris: OECD), 93–105.
- Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, H., Graesser, A. C., et al. (2014). Domain-general problem solving skills and education in the 21st century. *Educ. Res. Rev.* 13, 74–83. doi: 10.1016/j.edurev.2014.10.002
- Greiff, S., Wüstenberg, S., and Funke, J. (2012). Dynamic problem solving: a new assessment perspective. *Appl. Psychol. Meas.* 36, 189–213. doi: 10.1177/0146621612439620
- Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F., and Funke, J. (2013a). Computer-based assessment of complex problem solving: concept, implementation, and application. *Educ. Technol. Res. Dev.* 61, 407–421. doi: 10.1007/s11423-013-9301-x
- Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., and Csapó, B. (2013b). Complex problem solving in educational contexts – something beyond g: concept, assessment, measurement invariance, and construct validity. *J. Educ. Psychol.* 105, 364–379.
- Griffin, P., and Care, E. (eds.) (2015). *Assessment and Teaching of 21st Century Skills. Methods and Approach*. Dordrecht: Springer Netherlands.
- Griffin, P., McGaw, B., and Care, E. (eds.) (2012). *Assessment and Teaching of 21st Century Skills*. New York, NY: Springer.
- Hutchings, P., Kinzie, J., and Kuh, G. D. (2015). "Evidence of student learning. What counts and what matters for improvement," in *Using Evidence of Student Learning to Improve Higher Education*, eds D. Kuh, S. O. Ikenberry, N. A. Jankowski, T. R. Cain, P. T. Ewell, P. H. Hutchings, and J. Kinzie (San Francisco, CA: Jossey-Bass), 1–23.
- Ikenberry, S. O., and Kuh, G. D. (2015). "From compliance to ownership: why and how colleges and universities assess student learning," in *Using Evidence of Student Learning to Improve Higher Education*, eds D. Kuh, S. O. Ikenberry, N. A. Jankowski, T. R. Cain, P. T. Ewell, P. H. Hutchings, and J. Kinzie (San Francisco, CA: Jossey-Bass), 27–50.
- Kong, S. C., Chan, T. W., Griffin, P., Hoppe, U., Huang, R., Kinshuk, L., et al. (2014). E-learning in school education in the coming 10 years for developing 21st century skills: critical research issues and policy implications. *Educ. Technol. Soc.* 17, 70–78.
- Lopez, E. J., Shavelson, R. J., Nandagopal, K., Szu, E., and Penn, J. (2014). Factors contributing to problem-solving performance in first-semester organic chemistry. *J. Chem. Educ.* 91, 976–981. doi: 10.1021/ed400696c
- Molnár, G., Greiff, S., and Csapó, B. (2013). Inductive reasoning, domain specific and complex problem solving: relations and development. *Think. Skills Creat.* 9, 35–45. doi: 10.1016/j.tsc.2013.03.002
- Molnár, G., Greiff, S., Wüstenberg, S., and Fischer, A. (2017). "Empirical study of computer-based assessment of domain-general complex problem-solving skills," in *The Nature of Problem Solving. Using Research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (OECD: Paris), 125–140.
- Muthén, L. K., and Muthén, B. O. (2012). *Mplus User's Guide, 7th Edn.* Los Angeles, CA: Muthén & Muthén.
- National Research Council (2013). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington, DC: National Academies Press.
- Neubert, J. C., Mainert, J., Kretschmar, A., and Greiff, S. (2015). The assessment of 21st century skills in industrial and organizational psychology: complex and collaborative problem solving. *Ind. Organ. Psychol.* 8, 238–268. doi: 10.1017/iop.2015.14
- OECD (2004). *Learning for Tomorrow's World: First Results from PISA 2003*. Paris: OECD.
- OECD (2013a). *PISA 2012 Assessment and Analytic Framework*. Paris: OECD
- OECD (2013b). *PISA 2015 Draft Collaborative Problem Solving Framework*. Paris: OECD.
- OECD (2014). *PISA 2012 Results: Creative Problem Solving. Students' Skills in Tackling Real-Life Problems (Vol. V)*. Paris: OECD.
- OECD (2016). *PISA 2015 Results: Excellence and Equity in Education (Vol. I)*. Paris: OECD.
- Qian, M., and Clark, K. R. (2016). Game-based learning and 21st century skills: a review of recent research. *Comput. Hum. Behav.* 63, 50–58. doi: 10.1016/j.chb.2016.05.023
- Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., et al. (2012). The genetics lab. Acceptance and psychometric characteristics of a computer-based microworld to assess complex problem solving. *Psychol. Test Assess. Model.* 54, 54–72.
- Tóth, K., Rölke, H., Goldhammer, F., and Barkow, I. (2017). "Educational process mining: New possibilities for understanding students' problem-solving skills," in *The Nature of Problem Solving. Using Research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (Paris: OECD), 193–210.
- Tremblay, K., Lalancette, D., and Roseveare, D. (2012). *Assessment of Higher Education Learning Outcomes (AHELO). Feasibility Study Report Vol. 1. Design and Implementation*. Paris: OECD.
- Trilling, B., and Fadel, C. (2009). *21st Century Skills: Learning for Life in Our Times*. San Francisco, CA: John Wiley & Sons.
- Vainikainen, M. P., Wüstenberg, S., Kupiainen, S., Hotulainen, R., and Hautamäki, J. (2015). Development of learning to learn skills in primary school. *Int. J. Lifelong Educ.* 34, 376–392. doi: 10.1080/02601370.2015.1060025
- Wiley, E. W., Shavelson, R. J., and Kurpius, A. A. (2014). On the factorial structure of the SAT and implications for next-generation college readiness assessments. *Educ. Psychol. Meas.* 74, 859–874. doi: 10.1177/0013164414528332
- Wirth, J., and Klieme, E. (2003). Computer-based assessment of problem solving competence. *Asses. Educ.* 10, 329–345. doi: 10.1080/0969594032000148172
- Wu, M., Adams, M. J., and Wilson, M. (1998). *ACER ConQuest [Computer Program and Manual]*. Hawthorn, Australia: ACER Press.
- Wüstenberg, S., Greiff, S., and Funke, J. (2012). Complex problem solving – more than reasoning? *Intelligence* 40, 1–14. doi: 10.1016/j.intell.2011.11.003
- Yip, M. C. W. (2013). Learning strategies and their relationships to academic performance of high school students in Hong Kong. *Educ. Psychol.* 33, 817–827. doi: 10.1080/01443410.2013.794493
- Zlatkin-Troitschanskaia, O., Shavelson, R. J., and Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Stud. High. Educ.* 40, 393–411. doi: 10.1080/03075079.2015.1004241

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer, CL, and handling Editor declared their shared affiliation.

Copyright © 2017 Csapó and Molnár. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# The Association between Motivation, Affect, and Self-regulated Learning When Solving Problems

Martine Baars<sup>1\*</sup>, Lisette Wijnia<sup>1,2</sup> and Fred Paas<sup>1,3</sup>

<sup>1</sup> Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Rotterdam, Netherlands,

<sup>2</sup> Roosevelt Center for Excellence in Education, HZ University of Applied Sciences, Middelburg, Netherlands, <sup>3</sup> Early Start Research Institute, University of Wollongong, Wollongong, NSW, Australia

## OPEN ACCESS

### Edited by:

Wolfgang Schoppek,  
University of Bayreuth, Germany

### Reviewed by:

Maria Tulis,  
University of Augsburg, Germany  
Rakefet Ackerman,  
Technion – Israel Institute  
of Technology, Israel

Yael Sidi contributed to the review of  
Rakefet Ackerman

### \*Correspondence:

Martine Baars  
baars@fsw.eur.nl

### Specialty section:

This article was submitted to  
Cognitive Science,  
a section of the journal  
Frontiers in Psychology

**Received:** 14 March 2017

**Accepted:** 24 July 2017

**Published:** 08 August 2017

### Citation:

Baars M, Wijnia L and Paas F (2017)  
The Association between Motivation,  
Affect, and Self-regulated Learning  
When Solving Problems.  
Front. Psychol. 8:1346.  
doi: 10.3389/fpsyg.2017.01346

Self-regulated learning (SRL) skills are essential for learning during school years, particularly in complex problem-solving domains, such as biology and math. Although a lot of studies have focused on the cognitive resources that are needed for learning to solve problems in a self-regulated way, affective and motivational resources have received much less research attention. The current study investigated the relation between affect (i.e., Positive Affect and Negative Affect Scale), motivation (i.e., autonomous and controlled motivation), mental effort, SRL skills, and problem-solving performance when learning to solve biology problems in a self-regulated online learning environment. In the learning phase, secondary education students studied video-modeling examples of how to solve hereditary problems, solved hereditary problems which they chose themselves from a set of problems with different complexity levels (i.e., five levels). In the posttest, students solved hereditary problems, self-assessed their performance, and chose a next problem from the set of problems but did not solve these problems. The results from this study showed that negative affect, inaccurate self-assessments during the posttest, and higher perceptions of mental effort during the posttest were negatively associated with problem-solving performance after learning in a self-regulated way.

**Keywords:** affect, motivation, mental effort, self-regulated learning, problem-solving performance

## INTRODUCTION

Problem-solving is an important cognitive process, be it in everyday life, at work or at school. Problem-solving is the process in which people put effort into closing the gap between an initial or current state (also called givens) and the goal state (Mayer, 1992; Jonassen, 2011; Schunk, 2014). Research has shown that self-regulated learning (SRL) skills are important for effective problem-solving (e.g., Ackerman and Thompson, 2015). Self-regulated learning can be defined as “the degree to which learners are metacognitively, motivationally, and behaviorally active participants in their own learning process” (Zimmerman, 2008, p. 167). Not surprisingly, SRL skills like monitoring and regulating learning processes are important for learning during school years and in working life (Winne and Hadwin, 1998; Zimmerman, 2008; Bjork et al., 2013; for a meta-analysis see, Dent and Koenka, 2016). The process by which learners use SRL skills such as monitoring and control in reasoning tasks, problem-solving, and decision-making processes is also called meta-reasoning (Ackerman and Thompson, 2015). Monitoring judgments about problem-solving tasks

and decision-making processes could be related to the effort learners put into finding and using different types of strategies to solve the problem or make a decision.

Self-regulated learning skills are especially important in learner-controlled, online learning environments in which students need to be able to accurately keep track of their own learning process (i.e., monitoring) and have to make complex decisions about what problem-solving task to choose next during their learning process (i.e., regulation choices). Apart from the high cognitive demands imposed by SRL, which have been investigated frequently in previous research (e.g., Dunlosky and Thiede, 2004; Griffin et al., 2008; Van Gog et al., 2011), learning to solve problems in a self-regulated way also imposes demands on affective and motivational resources (Winne and Hadwin, 1998; Pekrun et al., 2002; Sperling et al., 2005; Efklides, 2011; Mega et al., 2014). The current study investigated the role of affect and motivation in learning problem-solving tasks in a complex learner-controlled online learning environment for secondary education students.

## Learning to Solve Problems

There are many different kinds of problem-solving tasks, varying from well-structured transformation problems that have a clearly defined goal and solution procedure, to ill-structured problems that do not have a well-defined goal or solution procedure (Jonassen, 2011). In educational settings like schools, universities, or trainings, students usually solve well-structured problems, especially in the domains of science, technology, engineering, and mathematics (STEM domains). Although, well-structured problems, such as math and biology problems encountered in primary and secondary education, can typically be solved by applying a limited and known set of concepts, rules, and principles, they are considered complex in terms of the high number of interacting elements that needs to be considered simultaneously in working memory (WM) during the problem-solving process (e.g., Kalyuga and Singh, 2016).

For learning to solve such complex problems, it is efficient to “borrow” and “reorganize” knowledge of others (Sweller and Sweller, 2006) by learning from examples, such as worked examples and modeling examples (Van Gog and Rummel, 2010). A worked example is a step-by-step worked-out solution to a problem-solving task that students can study. Research in the context of cognitive load theory (CLT; Paas et al., 2003a; Sweller et al., 2011) has shown that for novices, studying worked examples of how the problem should be solved, is a more effective strategy for learning to solve problems than solving equivalent conventional problems (i.e., the worked example effect; Sweller and Cooper, 1985; Paas, 1992; for reviews see, Sweller et al., 1998). According to CLT, having learners study worked examples is an effective way to reduce the extraneous load that is imposed by conventional problem-solving, because the learner can devote all available WM capacity to studying the worked-out solution and constructing a schema for solving such problems in long-term memory (Paas and Van Gog, 2006). In a modeling example, an adult or peer model performing a task can be observed, either face to face, on

video, via a screen recording made by the modeling person, or as an animation (Van Gog and Rummel, 2010). According to social-cognitive theory (Bandura, 1986), the learner can construct a mental representation of the task that is being modeled, and use it to perform the task at a later point in time.

According to the resource-allocation framework by Kanfer and Ackerman (1989) and CLT (Sweller et al., 1998) it can be assumed that the competition for WM resources between learning to solve a problem and self-regulation processes can have negative effects on either or both of these processes. For example, a student working on a complex problem-solving task needs most cognitive resources to perform the task itself, which leaves little resources to monitor and regulate learning. During the learning process, it could therefore be beneficial to study worked examples. Studying the step-by-step explanation on how to solve the problem leaves more WM resources for the construction of cognitive schemas (i.e., learning) than solving problems (i.e., worked example effect; Sweller, 1988; for reviews see, Sweller et al., 1998; Van Gog and Rummel, 2010). Therefore, it can be expected that the surplus cognitive capacity that becomes available by the reduction of extraneous cognitive load can be devoted to activities that further contribute to learning performance, such as self-regulation processes (Paas and Van Gog, 2006).

Despite this expectation, SRL skills, such as monitoring one's own learning processes, have been found to be suboptimal when studying worked examples (Baars et al., 2014a,b, 2016). A possible reason for this finding is that students' monitoring process when learning from worked examples can be prone to an illusion of competence. Students overestimate their competence to solve a problem when information about the problem solution is present during studying (Koriat and Bjork, 2005, 2006; Bjork et al., 2013). Similarly, studies with primary and secondary education students have found that students who learned to solve problems by studying worked examples showed inaccurate monitoring performance, because they overestimated their future test performance (Baars et al., 2013, 2014a,b, 2016; García et al., 2015). Yet, accurate monitoring is a prerequisite for effective self-regulation (cf. Thiede et al., 2003), and plays an important role in learning to solve problems (Mayer, 1992; Zimmerman and Campillo, 2003).

In a previous study by Kostons et al. (2012) video models were used to explain to secondary education students how to solve hereditary problems and additionally used the video-modeling examples to train students to self-assess their performance and make regulation choices in a learner-controlled environment. In the study, problem-solving performance, self-assessment, and task selection accuracy improved. These results are promising. However, large standard deviations in self-assessment accuracy and task selection were found, suggesting large individual differences in these SRL skills (Kostons et al., 2012), indicating that some students benefitted more from the video-modeling examples than others. Among others, Kostons et al. (2012) have suggested that these differences might be explained by motivation and affect.

## Problem-Solving, Affect, and Motivation

Students' affect and motivation can facilitate or hinder students when learning to solve problems in a self-regulated way. Affect was found to influence the use of different strategies (e.g., organization of study time, summarizing materials), SRL activities (e.g., reflecting on learning), and motivation; all factors that can impact academic achievement (Pekrun et al., 2002; Efklides, 2011; Mega et al., 2014). Moreover, in the domain of problem-solving, positive and negative affect were found to influence the problem-solving strategies (e.g., seeking and use of information) that students used (Spering et al., 2005).

According to theories on SRL both affect and motivation play an important role in SRL (e.g., Winne and Hadwin, 1998; Pintrich, 2004; Efklides, 2011). According to Efklides (2011), the interaction between metacognition, motivation, and affect is the basis of students' SRL. In Efklides' Metacognitive and Affective model of SRL (MASRL model), SRL is not only determined by a person's goal, but also by an interaction between metacognitive experiences, motivation, and affect during task performance. In line with the MASRL model, a study by Mega et al. (2014) showed that both negative and positive affect influence different aspects of SRL. For instance, positive affect was positively related to the evaluation of learning performance and metacognitive reflection during studying. In addition, both negative and positive affect were also shown to influence students' motivation. For example, positive affect enhanced students' beliefs on incremental theory of intelligence and their academic self-efficacy. Positive affect was found to have a greater impact on both SRL abilities and motivation compared to negative affect. SRL abilities and motivation in turn were predictive of academic achievement. However, the effect of motivation on academic achievement was larger than the effect of SRL abilities on academic achievement. Mega et al. (2014) further showed that the relation between affect and academic achievement was mediated by motivation and SRL abilities.

Although the study by Mega et al. (2014) showed the influence of affect on motivation and SRL abilities and subsequent academic performance, the implications for learning a variety of subjects during school years are still not clear. In the study by Mega et al. (2014), two general academic achievement indicators were used with undergraduate students from different disciplines. These general indicators of academic achievement were productivity (i.e., number of exams passed) and ability (i.e., GPA). These indicators are domain general and therefore, it is unclear whether these results would also apply to the domain of problem-solving or to task-specific performance within a domain.

### The Role of Affect in Problem-Solving

In the domains of problem-solving and decision-making, it was found that positive affect facilitates flexible and creative thinking, and decision-making in complex environments such as medical decision-making (Fiedler, 2001; Isen, 2001). In a review by Isen (2001), it was shown that if the situation is important or interesting to a person, positive affect will enhance systematic, cognitive processing and thereby make this process more efficient and innovative. Positive affect was found to improve generosity,

creativity, variety seeking, negotiation, and decision-making in a range of different domain and contexts such as problem-solving (e.g., Duncker's problem), consumer decision-making, coping with stressful life-events, bargaining when buying and selling appliances, car choice, and medical diagnosis. For example, physicians with positive affect induced by a small gift (i.e., a box of candy), scored higher on creativity as measured by the Remote Associates Test (Estrada et al., 1994). Also, in a study by Politis and Houtz (2015) it was found that middle school students who watched a positive video program to induce positive affect generated a greater number of ideas compared to students who watched a neutral video program. More closely related to problem-solving tasks that can be solved in a stepwise manner, Brand et al. (2007) demonstrated that affect influenced solving the Tower of Hanoi (ToH) problem in adult students. After inducing negative affect, participants needed more repetitions to learn to solve the ToH problem and performed worse on the transfer tasks compared to participants with an induced positive mood.

In contrast to the findings showing that positive affect can facilitate problem-solving performance (e.g., Isen, 2001), some studies found that positive affect does not facilitate problems solving. In a study by Kaufmann and Vosburg (1997) high school students rated their affect at the beginning of the experiment and then engaged in solving insight problems which were unstructured and high in novelty and analytical tasks from an intelligence test. It was found that positive affect reduced problem-solving performance on the insight problems but not on the analytical tasks. These results were replicated in a second experiment with college students whose affect was induced using positive, negative, and neutral videotapes. The authors suggest that because students in their study did not receive any feedback and had to judge their solution for themselves, students with positive affect probably stopped searching for task-relevant information earlier than students with a negative mood. In line with this hypothesis, Spering et al. (2005) found that negative affect led to more detailed information search during complex problem-solving. In the study by Spering et al. (2005) with 74 undergraduate and graduate students, positive and negative affect were induced and the effect on complex problem-solving (CPS) was investigated. In CPS the situation is complex, variables are connected, there is a dynamic development of the situation, the situation is non-transparent, and people can pursue multiple goals (Funke, 2001). Positive and negative affect were induced by positive and negative performance feedback (Spering et al., 2005). Although, positive or negative affect increased as was intended, the results showed that positive and negative affect did not influence performance (Spering et al., 2005). However, negative affect did lead to more detailed information search and a more systematic approach (Spering et al., 2005).

To sum up, positive affect could facilitate problem-solving and decision-making. Yet, this seems to be dependent on the type of problems used in the different studies. The problem-solving tasks in the review by Isen (2001) were more structured or transparent than the ones used in the studies by Kaufmann and Vosburg (1997) and by Spering et al. (2005). For more structured problems, positive affect could facilitate problem-solving. If applied on learning to solve well-structured, stepwise



hereditary problems in secondary education, one would expect positive affect to facilitate self-regulation of the learning process and problem-solving performance. The role of motivation, as described in the MASRL model by Efklides (2011) could interact with this process.

### The Role of Motivation

Self-determination theory (SDT; Deci and Ryan, 2000; Ryan and Deci, 2000a,b) predicts that students use more effort and process the materials more deeply when they find the learning materials interesting. There are several types of motivation which can be placed on a continuum of the degree of experienced autonomy. Students with a high degree of autonomous motivation experience volition and psychological freedom. They study because the subject is interesting to them or it brings them satisfaction (i.e., intrinsic motivation). Also, doing the task could be valuable for attaining personal goals or development (i.e., identified motivation). However, students who score high on controlled motivation experience a low degree of autonomy and experience pressure. This pressure can come from within the student (i.e., introjected motivation). For example, students feel pressure to avoid feelings of shame, or pressure can come from an external source, such as demands from a teacher or a parent (i.e., external motivation).

Autonomous motivation types are associated with better learning outcomes, persistence, and psychological well-being relative to controlled motivation types. Autonomous motivation types were found to be related to better text comprehension (e.g., Vansteenkiste et al., 2004) and self-reported academic achievement (e.g., Vansteenkiste et al., 2009). Furthermore, motivation based on interest has been associated with better problem-solving performance (for a review see Mayer, 1998) and better SRL abilities such as effort regulation (i.e., controlling effort and attention) and metacognitive strategy use (i.e., checking and correcting one's own learning behavior; Vansteenkiste et al., 2009). Moreover, it was found that students who indicated higher levels of interest for a course (i.e., an autonomous reason for studying), were more likely to use strategies to monitor and regulate their learning (Pintrich, 1999).

In summary, next to enhancing learning and problem-solving performance, autonomous motivation could also facilitate the use of SRL skills during learning. Furthermore, in multiple studies by Pekrun et al. (2002) intrinsic motivation was found to be related to positive affect such as enjoyment, hope, and pride. Also, negative affect such as boredom and hopelessness were found to be negatively related to intrinsic motivation and effort.

### Present Study and Hypotheses

The relation between affect, self-assessment accuracy, making complex decisions about the learning process (i.e., regulation choice complexity), perceived mental effort and motivation was investigated in a learner-controlled, online environment, in which students could monitor and regulate their own learning. In this environment students first received video-modeling examples teaching them how to solve stepwise, hereditary problem-solving tasks, how to make a self-assessment (i.e., monitoring), and how to select the next task (i.e., regulation

choice). In each video-modeling example, after solving the problem, the model rated the perceived amount of invested mental effort (Paas, 1992), made a self-assessment of his/her performance over the five steps, made a regulation choice, and explained these actions (cf., Kostons et al., 2012; Raaijmakers et al., unpublished). After the video-modeling examples, students were asked to select and practice four problems from an overview with 75 problem-solving tasks. Affect was measured at the start of the study. Mental effort, self-assessment accuracy, and regulation choice complexity were measured during the posttest. Motivation was measured at the end of study.

Although the problems in the learning phase were well-structured, the online learning environment in which students had to learn to solve them could be considered a complex problem-solving environment that required cognitive activities such as monitoring and planning with problem-solving tasks of different complexity levels (Osman, 2010). That is, during the learning phase students had to choose the problem-solving task they wanted to work on next from a task database with 75 tasks arranged by five complexity levels (see **Figure 1**). Task complexity of the well-structured problems was defined in terms of element interactivity: the higher the number of interacting information elements that a learner has to relate and keep active in WM when performing a task, the higher the complexity of that task and the higher the cognitive load it imposes (Sweller et al., 1998; Sweller, 2010). The easier problems consisted of less interacting information elements (e.g., two generations, one unknown, and deductive reasoning) compared to the more difficult problems (e.g., three generations, two unknowns, and both deductive and inductive reasoning). In addition, monitoring the learning process and choosing the next task at a certain complexity level based on monitoring processes also adds to the complexity of the learning process and imposes cognitive load upon the learner (e.g., Griffin et al., 2008; Van Gog et al., 2011). Taken together, monitoring learning and choosing tasks with different levels of interacting elements, created a complex problem-solving environment in which the current study took place.

We expected positive and negative affect, self-assessment accuracy, regulation choice complexity, perceived mental effort, and autonomous and controlled motivation to be predictors of problem-solving performance. More specifically, we expected positive affect measured at the beginning of the study to be a positive predictor of problem-solving performance (cf., Isen, 2001, Hypothesis 1a), whereas negative affect measured at the beginning of the study was expected to be a negative predictor of problem-solving performance (Hypothesis 1b).

According to theories of SRL (e.g., Winne and Hadwin, 1998; Zimmerman, 2008), we expected self-assessment accuracy during the posttest to be positively associated with problem-solving performance at the posttest (Hypothesis 2a). We further hypothesized that regulation choice complexity during the posttest would be positively associated with problem-solving performance at the posttest (Hypothesis 2b). Based on theories of SRL one would expect students to make regulation choices based on monitoring processes. Therefore, the more complex students' regulation choices were, the better they think they

Complexity level	Support level	Learning tasks				
Complexity 5: -3 generations -2 unknowns -Multiple answers -Both ways	No support	Eye color	Hair structure	Milk allergy	Depression	Huntington
	Low support	Eye color	Hair structure	Dog tail length	Wolfram	Flower color
	High support	Eye color	Hair structure	Cystic Fibrosis	Fruit flies	Rat fur
Complexity 4: -3 generations -1 unknown -Multiple answers -Both ways	No support	Eye color	Hair structure	Albinism	Cat fur shape	Fruit flies
	Low support	Eye color	Hair structure	Fruit flies	Tongue curling	Flower color
	High support	Eye color	Hair structure	Pea plant	Dimples	Depression
Complexity 3: -2 generations -1 unknown -Multiple answers -Inductive	No support	Eye color	Hair structure	Fruit flies	Chicken beak	Wolfram
	Low support	Eye color	Hair structure	Dog tail length	Apple tree	Milk allergy
	High support	Eye color	Hair structure	Freckles	Flower color	Earlobes
Complexity 2: -2 generations -1 unknown -Multiple answers -Deductive	No support	Eye color	Hair structure	Flower color	Widow's peak	P.R.A.
	Low support	Eye color	Hair structure	Cat fur shape	Albinism	Pea plant
	High support	Eye color	Hair structure	Tongue curling	Apple tree	Fruit flies
Complexity 1: -2 generations -1 unknown -Single answer -Deductive	No support	Eye color	Hair structure	Cat fur shape	Apple tree	Depression
	Low support	Eye color	Hair structure	Sickle cell	Chicken beak	Guinea pigs
	High support	Eye color	Hair structure	Huntington	Milk allergy	Cleft lip

**FIGURE 1 |** Task database containing the 75 problem-solving tasks showing the different levels of complexity, different levels of support, and the different surface features of the learning tasks (Raaijmakers et al., unpublished).

performed (assuming that monitoring and regulation processes would approach actual performance and are more or less accurate).

Competition for WM resources between learning to solve a problem and self-regulation processes can have negative effects on either or both of these processes (Kanfer and Ackerman, 1989; Sweller et al., 1998). Based on the efficiency account of Paas and Van Merriënboer (1993; see also Van Gog and Paas, 2008) we assumed that the combination of perceived mental effort during the posttest and posttest performance would be indicative for the quality of learning (i.e., problem-solving) during the learning phase. Therefore, we hypothesized that students who managed to gain more knowledge during the learning phase, would experience lower mental effort during the posttest and obtain higher posttest performance than students who experience higher mental effort during the posttest. Therefore, perceived mental effort during the posttest was expected to be a negative predictor of problem-solving performance (Hypothesis 3a) and show a negative relation with SRL skills such as monitoring (Hypothesis 3b) and regulation choices (Hypothesis 3c) as measured during the posttest.

According to SDT, autonomous motivation is associated with better learning outcomes and SRL when compared to controlled motivation (Deci and Ryan, 2000). In line with the findings by Vansteenkiste et al. (2009), we expected autonomous motivation to be positively related to problem-solving performance (Hypothesis 4a), whereas controlled motivation was expected

to be negatively related to problem-solving performance (Hypothesis 4b).

## MATERIALS AND METHODS

### Participants

Participants were 136 secondary school students ( $M_{\text{age}} = 13.73$ ,  $SD = 0.58$ , 74 girls) from the second year in the higher education track. All students gave their consent to participate in this study. Students' parents received a letter in which information about the study was provided and parents were asked for their consent.

### Materials

Students participated in the computer rooms at their schools. They entered an online learning environment<sup>1</sup> of which the content was created by the researchers for the purpose of this study. All measures were assessed online.

### Affect Questionnaire

At the beginning of the study, all students filled out the 20-item Positive Affect and Negative Affect Scale (i.e., PANAS) on a 5-point scale (Watson et al., 1988). For both the positive affect scale (10-items) and the negative affect scale (10-items) an average score was calculated per participant. The reliability for the

<sup>1</sup>www.qualtrics.com

positive affect scale measured with Cronbach's alpha was  $\alpha = 0.76$  and  $\alpha = 0.76$  for the negative affect scale.

### Pretest and Posttest

The pretest and posttest consisted of three well-structured problem-solving tasks about hereditary problems based on the laws of Mendel which differed in complexity in terms of element interactivity (cf. Kostons et al., 2012). All problem-solving tasks consisted of five steps: (1) determining genotypes from phenotypes, (2) constructing a family tree, (3) determining whether the reasoning should be deductive or inductive, (4) filling out the crosstabs, (5) distracting the answer from the crosstabs (see Appendix A for an example). Problem-solving tasks 1 and 2 could both be solved by deducting the genotype of the child based on information about the parents. Task 2 was more difficult because the genotype of the parents was heterozygote vs. homozygote in task 1, which means that more interacting information elements needed to be taken into account during the problem-solving process. Problem-solving task 3 was the most complex problem-solving task because the genotype of one of the parents had to be induced based on information about the other parent and the child (i.e., inductive). This added more interactive information elements, and therefore complexity to the problem-solving process. The pretest and posttest were isomorphic to each other (i.e., different surface features were used). On both tests, students could score 1-point per correctly solved step adding up to 5-points per problem-solving tasks and 15-points in total.

### Video-Modeling Examples

Two video-modeling examples showed how to solve a hereditary problem step by step. The hereditary problems explained in the videos had a similar solution procedure because in both videos the goal was to find the genotype of the child based on information about the parents (i.e., deductive). The surface features were different between the problems explained in the videos (i.e., nose bridge and tongue folding). In the videos, a model was thinking aloud about how to solve the problem and wrote down the solution step by step. One video had a female model and the other video had a male model explaining how to solve a problem (see Appendix B for an example). In each video after solving the problem, the model rated their mental effort on a 9-point scale (Paas, 1992), made a self-assessment of their performance over the five steps, made a regulation choice, and explained these actions (cf. Raaijmakers et al., unpublished). The regulation choice was based on a heuristic which uses performance and effort to choose the next task. The heuristic states that when one has a high performance combined with low mental effort one needs to choose a more difficult task, whereas with low performance and high effort one should choose an easier task (see Paas and Van Merriënboer, 1993; Van Gog and Paas, 2008).

### Mental Effort Rating

After each posttest question, mental effort invested in solving the posttest problems was measured by asking: 'How much effort did you invest in solving this problem?' Students could respond on a

9-point scale, ranging from 1 (*very, very low mental effort*) to 9 (*very, very high mental effort*, Paas, 1992; Paas et al., 2003b; Van Gog et al., 2012). The mean mental effort rating for the pretest and the posttest was calculated. Unfortunately, six students did not fill out all the mental effort ratings and were left out of the analysis of the mental effort data ( $n = 130$ ).

### Self-assessment

Students made a self-assessment of their performance as a measure of self-monitoring after each posttest problem-solving task (cf. Baars et al., 2014a). Students rated which steps of the problem they thought they had solved correctly (0 indicating every step was wrong and 5 indicating every step was correct). Self-assessment accuracy was measured as absolute deviation (Schraw, 2009). Thus, absolute accuracy was calculated as the square root of the squared difference between actual performance and rated self-assessment per problem-solving task. The lower absolute deviation is, the smaller the distance between the self-assessment and the actual performance is and therefore, the more accurate self-monitoring (i.e., self-assessment) was. Unfortunately, six students did not fill out all the self-assessments and were left out of the analysis of the self-assessment data ( $n = 130$ ).

### Regulation Choice Complexity

During the posttest, the complexity of the regulation choices of students was measured. Students could choose problem-solving tasks to study next from a database with 75 problem-solving tasks at five complexity levels (see Figure 1, cf. Kostons et al., 2012; Raaijmakers et al., unpublished). They choose a task after solving each of the three posttest problems. The complexity of the regulation choice was measured with 1 being the easiest task to choose and 5 being the most difficult task to choose. The simplest problems consisted of 2 generations, 1 unknown, single answer, and deductive solution procedures. The most complex problems consisted of 3 generations, 2 unknowns, multiple answers, and deductive and inductive solutions procedures (for an overview see Figure 1). The level of support was not included in the level of complexity. Note, during the posttest students did not actually study the tasks they choose and they were made aware of that. The mean regulation choice complexity score for the posttest was calculated. There were 33 students who did not make a regulation choice and therefore they were left out of the analysis of regulation choice data ( $n = 103$ ).

### Motivation Questionnaire

At the end of the study, all students filled out a 16-item task-specific version of the academic self-regulation scale (Vansteenkiste et al., 2004). In four subscales, they had to indicate why they worked on solving the hereditary problem-solving tasks: (1) external (e.g., "... because I am supposed to do so"), (2) introjected (e.g., "... because I would feel guilty if I did not do it"), (3) identified (e.g., "... because I could learn something from it"), and (4) intrinsic motivation (e.g., "... because I found it interesting"). Items were measured on a 5-point Likert-type scale ranging from 1 (*not at all true*) to 5 (*totally true*). The

four subscales were combined into an autonomous motivation composite (intrinsic and identified motivation) and a controlled motivation composite (introjected and external motivation; cf. Vansteenkiste et al., 2004). There were 10 students who did not complete the motivation questionnaire and therefore they were left out of the analysis of the motivation data. For the autonomous motivation composite ( $n = 126$ ) Cronbach's alpha was  $\alpha = 0.89$ . For the controlled motivation composite ( $n = 126$ ) Cronbach's alpha was  $\alpha = 0.65$ .

## Procedure

In 50-min sessions in the computer room at their schools, students participated in the current study using an online learning environment<sup>2</sup>. In **Figure 2**, the procedure of the study is depicted. First, all students filled out the affect questionnaire. Then they took the pretest which was followed by two video-modeling examples. Then students entered the SRL phase in which they practiced with four problem-solving tasks of their choice from a database with 75 problem-solving tasks at five complexity levels (see the database in **Figure 1**). Students also practiced with rating their perceived mental effort, self-assessment, and regulation choices. Then after practicing four problem-solving tasks, students took a posttest with three problem-solving tasks of different complexity. Students' perceived mental effort, self-assessments, regulation choices, and problem-solving performance were measured. Finally, all students filled out the motivation questionnaire.

## RESULTS

In **Table 1**, the descriptive statistics of the pretest, posttest, perceived mental effort, self-assessments during the posttest (raw score, bias, and absolute accuracy), positive and negative affect scale, and autonomous and controlled motivation can be found. In **Table 2**, the correlations between these variables are shown. Pretest performance was significantly positively related to posttest performance. Positive affect was significantly positively associated with negative affect, indicating that students who scored higher on positive feelings also scored higher

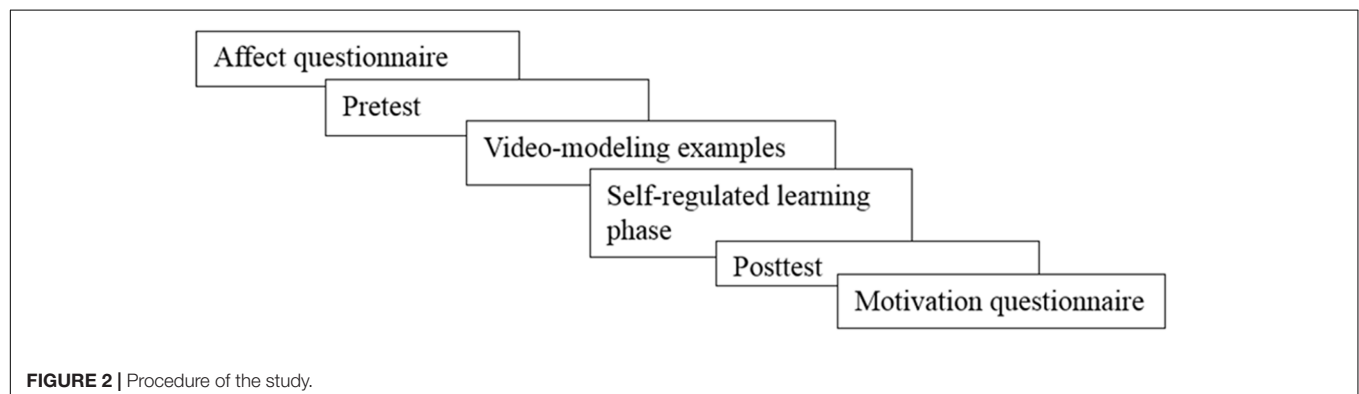
on negative feelings. Positive affect was significantly positively related to autonomous motivation. In line with Hypothesis 1b, negative affect was significantly negatively related to performance on the pretest and posttest, which indicated that students who reported more negative feelings scored lower on the tests.

In line with Hypothesis 2a, both self-assessment bias and absolute accuracy of self-assessments during the posttest were significantly negatively related to posttest performance. That is, the larger the difference between self-assessment and actual performance was, the lower posttest performance was. It seemed that students who are less accurate in their self-assessment also score lower on the posttest.

In support of Hypotheses 3b and c, the ratings of perceived mental effort showed a significant negative relation with the self-assessment raw score, bias, and complexity of regulation choices. This means that students who experienced a higher mental effort showed lower self-assessment and bias values, and choose less complex tasks to restudy. Both self-assessment raw scores and bias were positively correlated to the complexity of regulation choices. That is, the higher self-assessment raw scores and bias were, the more complex regulation choices were. In line with theories of SRL, this shows the sensitivity of regulation choices in relation to self-assessments (control sensitivity; Koriath and Goldsmith, 1996). Also, in line with Hypothesis 3a, perceived mental effort was significantly negatively related to posttest performance.

Autonomous motivation was significantly positively related to controlled motivation. It seems that students who scored higher on autonomous motivation also scored higher on the controlled motivation. Autonomous motivation also showed a significant negative relation with self-assessment absolute accuracy during the posttest. That is, students who scored higher on autonomous motivation had lower absolute accuracy scores which means that the deviation between their self-assessment and actual posttest performance was smaller. In other words, students with higher autonomous motivation also had more accurate self-assessments during the posttest. In support of Hypothesis 4b, autonomous motivation also showed a significant positive relation with posttest performance. This indicates that students who scored higher on autonomous motivation also scored higher on the posttest.

<sup>2</sup>www.qualtrics.com





**TABLE 1 |** Means and Standard Deviations for performance, affect, motivation, and self-regulation variables.

Variable (Range)	<i>n</i>	Mean (SD)
Pretest score (0–15)	136	2.05 (1.59)
Perceived mental effort pretest (1–9)	136	7.74 (1.88)
Posttest score (0–15)	136	4.82 (2.85)
Perceived mental effort posttest (1–9)	130	5.60 (2.33)
Self-assessment rating (0–5)	130	2.71 (1.46)
Self-assessment bias (–5 to +5)	130	1.11 (1.60)
Self-assessment absolute accuracy (0–5)	130	1.74 (1.14)
Regulation complexity (1–5)	103	2.30 (1.26)
Positive affect scale (1–5)	136	2.64 (0.64)
Negative affect scale (1–5)	136	1.42 (0.44)
Autonomous motivation scale (1–5)	126	2.25 (0.84)
Controlled motivation scale (1–5)	126	2.70 (0.62)

## Regulation Choices and Problem Complexity

The complexity level at which students selected a task for restudy and how they performed on the different complexity levels in the posttest were explored. Regulation choice complexity was not normally distributed. The mode of all three selection moments was regulation choice complexity 1. Therefore, a Friedman's ANOVA was conducted for regulation choice complexity at all three selection moments during the posttest. The regulation choice complexity differed significantly over the three moments,  $\chi^2(2) = 8.59$ ,  $p = 0.014$ . Wilcoxon tests were used to follow up this finding. It appeared that regulation choice complexity differed significantly between moments 1 (*Mean rank* = 2.14) and 3 (*Mean rank* = 1.83),  $T = 0.30$ ,  $r = 0.21$  (small effect size). Yet, no significant differences between selection moments 1 and 2 (*Mean rank* = 2.03) or 2 and 3 were found.

Furthermore, as a check on the complexity of the problem-solving tasks in terms of element interactivity, a repeated measures ANOVA with complexity levels as a within-subjects variable was performed. It showed that problem-solving performance on the posttest differed significantly between the

complexity levels of the problem-solving tasks,  $F(1,135) = 56.13$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.30$ . Performance on the least complex problem-solving task 1 ( $M = 2.45$ ,  $SD = 1.68$ ) was significantly higher compared to task 2 ( $M = 1.23$ ,  $SD = 1.40$ ),  $p < 0.001$ , and compared to task 3 ( $M = 1.14$ ,  $SD = 0.71$ ),  $p < 0.001$ . There was no significant difference between performance on tasks 2 and 3.

## Affect, SRL Skills, and Motivation As Predictors for Problem-Solving Performance

We performed stepwise regression with pretest performance in Model 1 and positive affect, negative affect, self-assessment accuracy during the posttest, regulation choice complexity, perceived mental effort, autonomous, and controlled motivation in Model 2. We assessed multicollinearity in accordance with the guidelines by Field (2009) by checking the VIF and tolerance values. The VIF provides an indication of whether a predictor has a strong relationship with the other predictor(s) and the tolerance statistic is defined as  $1/\text{VIF}$ . VIF values were well below 10 and tolerance was well above 0.2. Thus, collinearity was not a problem for our model (Field, 2009).

As shown in **Table 3**, Model 1 with pretest performance as a predictor of posttest problem-solving performance was significant,  $F(1,100) = 8.80$ ,  $p = 0.004$ ,  $R^2 = 0.08$ . Pretest performance was a significant positive predictor of posttest problem-solving performance.

In Model 2, positive affect, negative affect, posttest self-assessment accuracy, posttest regulation choice complexity, posttest perceived mental effort, autonomous and controlled motivation were added as predictors,  $F(8,93) = 4.89$ ,  $p < 0.001$ ,  $R^2 = 0.30$ . Model 2 explained more variance compared to Model 1,  $\Delta R^2 = 0.22$ ,  $p = 0.001$ . Pretest performance was again a significant positive predictor of posttest problem-solving performance in Model 2. In line with Hypothesis 1b, negative affect was a significant negative predictor of posttest problem-solving performance. That is, the more negative affect students reported, the lower their posttest performance was. Also, in support of Hypothesis 2a, self-assessment

**TABLE 2 |** Summary of intercorrelations between measures.

Measures	1	2	3	4	5	6	7	8	9	10	11
(1) Positive affect	—										
(2) Negative affect	0.20*	—									
(3) Pretest performance	–0.07	–0.26**	—								
(4) Autonomous motivation	0.32**	–0.05	0.13	—							
(5) Controlled motivation	0.12	0.02	0.02	0.23**	—						
(6) Mental effort (posttest)	–0.03	0.11	–0.01	–0.09	0.16	—					
(7) Self-assessment raw score	0.07	–0.01	0.07	0.12	–0.08	–0.53**	—				
(8) Self-assessment bias	0.03	0.16	–0.12	–0.07	–0.10	–0.32**	0.81**	—			
(9) Self-assessment absolute accuracy	–0.04	0.15	–0.11	–0.18*	–0.03	–0.12	0.63**	0.81**	—		
(10) Regulation choice complexity	–0.07	–0.00	–0.01	0.06	–0.12	–0.30**	0.27**	0.23*	0.16	—	
(11) Posttest performance	0.06	–0.22**	0.29**	0.30**	0.05	–0.27**	0.17	–0.44**	–0.40**	0.04	—

\*Correlation is significant at the 0.05 level (2-tailed). \*\*Correlation is significant at the 0.01 level (2-tailed). High values for self-assessment absolute accuracy indicate low accuracy.

**TABLE 3 |** Stepwise regression with predictors of problem-solving performance.

	<i>b</i>	<i>SE</i>	$\beta$	<i>p</i>
<b>Step 1</b>				
Constant	3.85	0.44		<0.001
Pretest performance	0.50	0.17	0.28	0.004
<b>Step 2</b>				
Constant	6.05	1.89		0.002
Pretest performance	0.34	0.16	0.19	0.035
Positive affect	0.49	0.42	0.11	0.252
Negative affect	-1.54	0.76	-0.19	0.046
Self-assessment accuracy	-0.83	0.24	-0.31	0.001
Regulation choice complexity	0.06	0.21	0.03	0.768
Mental effort posttest	-0.29	0.12	-0.24	0.013
Autonomous motivation	0.41	0.34	0.12	0.232
Controlled motivation	0.34	0.46	0.07	0.464

$R^2 = 0.08$  for Step 1;  $\Delta R^2 = 0.30$  for Step 2 ( $p = 0.001$ ).

accuracy was a significant negative predictor of posttest problem-solving performance. Self-assessment accuracy during the posttest was measured as absolute accuracy. The lower this measure is the more accurate self-assessments were. Thus, the negative relation with posttest performance means that the less accurate students' self-assessments were, the lower posttest performance was. Furthermore, in line with Hypothesis 3, perceived mental effort was a significant negative predictor of posttest problem-solving performance. That is, the higher perceived mental effort during the posttest was, the lower posttest problem-solving performance was.

## CONCLUSION AND DISCUSSION

The current study investigated the relation between affect (i.e., positive affect and negative affect), SRL skills (i.e., monitoring and regulation), perceived mental effort, motivation (i.e., autonomous and controlled motivation), and performance when learning to solve problems in a complex learner-controlled, online learning environment with secondary education students. Students performed worse on the more complex problems during the posttest. Also, regulation choice complexity was lower after the most difficult problem-solving task when compared to the least complex problem-solving task at the posttest. Interestingly, the results showed that students' negative affect, SRL skills, and perceived mental effort play a crucial role in learning to solve problems in a self-regulated way in a learner-controlled study environment.

In contrast to Hypothesis 1a, positive affect was not a significant predictor of problem-solving performance in the current study using well-structured problem-solving tasks with high element interactivity. This result does not fit previous findings showing that positive affect improves cognitive processing (e.g., Estrada et al., 1994; Isen, 2001; Brand et al., 2007; Politis and Houtz, 2015) and academic achievement (Mega et al., 2014). Possibly, this difference can be explained

by the way affect was measured and whether it was induced or not. Many of the studies reviewed by Isen (2001), in the studies by Estrada et al. (1994), Brand et al. (2007), and Politis and Houtz (2015) induced positive affect was found to improve different aspects of problem-solving performance. In the current study, positive affect was measured using a questionnaire at the beginning of the study. Therefore, it could be that positive affect measured by a rating provided by students does not have the same effect as induced positive affect on problem-solving performance. The effect of positive affect without inducement might be more prominent on more general measures of achievement made over a period of time (e.g., Mega et al., 2014). Interestingly, in the study by Kaufmann and Vosburg (1997) high school students also rated their affect and it was found that positive affect reduced performance on insight problems but not on analytical tasks. Yet, in our study we did not find a positive or negative association of positive affect with problem-solving performance on well-structured stepwise problems in high school. Furthermore, in our study students learned to solve problems in a self-regulated way and had to make decisions about which tasks to practice which made the learning process as a whole quite complex for students. Sperring et al. (2005) found that in CPS, performance was not affected by positive or negative affect which would be partially in line with our findings (i.e., no relation between positive affect and problem-solving performance). Yet, in line with earlier results (e.g., Pekrun et al., 2002; Mega et al., 2014) and our hypothesis, we found that negative affect influenced problem-solving performance. Specifically, in support of Hypothesis 1b, negative affect negatively predicted problem-solving performance. The difference between the results found by Kaufmann and Vosburg (1997) and the current study might be explained by the difference in the type of problem-solving tasks used in both studies. Although, element interactivity made the problems complex for students, the stepwise solving procedure also made the problem-solving tasks well-structured. Possibly, our problem-solving tasks were more transparent and therefore less complex than the insight problems used by Kaufmann and Vosburg (1997). Because of different dimension on which complexity can be defined (e.g., structure, element interactivity, and transparency), future research should investigate the relation of positive and negative affect with these different dimensions of complexity in problem-solving tasks.

In line with Hypothesis 2a and theories of SRL (e.g., Winne and Hadwin, 1998; Zimmerman, 2008), self-assessment accuracy was positively related to problem-solving performance. Students who were less accurate in their self-assessments, showed lower posttest problem-solving performance. Hence, monitoring seems an important prerequisite for successful learning to solve problems in a self-regulated way. However, there is a possibility that students who were high performers, were also better able to monitor their own learning. The results of the current study cannot establish the causality of this relation. Future research could use an experiment to investigate the effect of monitoring on problem-solving performance.

In contrast to Hypothesis 2b, the regulation choice complexity was not related to problem-solving performance. This might be explained by the way we operationalized regulation choices. That is, students had to choose what task they wanted to work on next. According to the discrepancy-reduction framework of regulation (e.g., Nelson et al., 1994) students would choose tasks in between their current state of learning and the goal state. Within this perspective on regulation of learning, choosing more difficult tasks would contribute to successful SRL. Yet, students might have chosen to select a task they were almost able to solve, which would be in line with the region of proximal learning to explain regulation of learning (e.g., Kornell and Metcalfe, 2006). Also, students might have chosen the task because they were curious about or just wanted to solve based on an agenda they might have had for themselves (i.e., agenda-based regulation, Ariel et al., 2009). For example, students could have been curious about the most complex problems or they wanted to finish as fast as possible and therefore choose the easiest problems. Also, regulation choices might have been inaccurate (i.e., deviate from actual performance). If students were not able to accurately monitor and/or regulate their own learning, regulation choice complexity would not be related to performance (cf. Baars et al., 2014a, 2016). This could also be caused by the fact that the regulation choices made during the posttest were not granted (i.e., students did not actually work on the problem they chose again). Future research could investigate the reasons students have to choose certain tasks to regulate their learning and if these choices are accurate in relation to their performance. In addition, future research could grant students their regulation choices and investigate how that would affect subsequent problem-solving performance.

Perceived mental effort during the posttest was significantly related to problem-solving performance which was in line with Hypothesis 3. That is, the more mental effort students experienced during the posttest, the lower their posttest performance was. This finding is in line with CLT (e.g., Sweller et al., 1998) and the efficiency account introduced by Paas and Van Merriënboer (1993; see also Van Gog and Paas, 2008). Yet, it would be interesting to follow up on this finding by including measures of perceived mental effort and performance during the learning phase in future research. That way the learning process and the relation to perceived mental effort could be investigated more elaborately. Furthermore, in the current study perceived mental effort was also related to the complexity of regulation choices during the posttest. That is, students who experienced higher mental effort during the posttest, chose less complex problems when making regulation choices during the posttest. Possibly, students used their perceived mental effort as an indicator to regulate their learning. This is in line with earlier research showing that students use their study effort to regulate their learning when regulation is data-driven (i.e., based on the ease of learning, Koriath et al., 2014). This seems sensible because mental effort was a significant predictor of problem-solving performance. This result provides support for studies showing that training students to use their perceived mental effort to regulate their learning when learning to solve problems can be effective (e.g., Kostons et al., 2012; Raaijmakers et al.,

unpublished). Future research could also include measures of perceived difficulty and self-efficacy to investigate the relation between perceived mental effort, task difficulty, self-efficacy, and performance during SRL.

Contrary to our expectations (Hypotheses 4a and 4b), motivation was not a significant predictor of problem-solving performance. Based on earlier studies (e.g., Vansteenkiste et al., 2009) autonomous motivation was expected to be a positive predictor of problem-solving performance. Yet, in the current study we did not investigate the different types of motivation (i.e., profiles: autonomous, identified, introjected, and external motivation, Vansteenkiste et al., 2009). Perhaps when taking into account the differences between motivational profiles, the effect of motivation on performance would be more pronounced. In addition, motivation was measured at the end of the experiment because students needed to be familiar with the materials used in the study. Yet, perhaps because of fatigue or boredom, students rated their motivation lower at the end of study compared to a measurement that would have been earlier on in the study. Future research could investigate this by placing the motivation questionnaire right after the pretest which would give students an idea of the materials without being mentally exhausted.

Limitations of the current study are the small number of secondary education students who could take part in the study. Future research could replicate the current study with more participants. This would also enable researchers to take into account different motivational profiles and their relation to positive and negative affect as predictors of problem-solving performance. Also, problem-solving performance was quite low and measured using a limited set of tasks during the posttest. It would be interesting to use more tasks for a longer period of time covering the SRL phase and posttest to investigate the effect of motivation and affect. We found positive and negative affect to be positively related which could have been caused by the intensity of affect (Diener et al., 1985). Future research could measure this dimension of affect to control for it. In addition, for both motivation and affect questionnaires were used as a measurement. The motivation questionnaire was task specific and therefore placed at the end of the study, which could have caused students to use their experience of success or failure during the posttest when filling out the motivation questionnaire. Future research could design experiments in which affect is induced and motivation is measured earlier during the study or through the learning behaviors of students.

In conclusion, the current study showed that negative affect, monitoring accuracy, and perceived mental effort are predictors of problem-solving performance of secondary education students learning to solve problems in a learner-controlled, online environment. The fact that these predictors were all negatively related to performance is an important indication that students need more support when learning to solve problems in a self-regulated way. Interventions to support SRL processes (e.g., training, cf. Kostons et al., 2012) and reduce mental effort involved in learning to solve problems (e.g., worked-examples, Sweller, 1988), could potentially prevent negative effects of

inaccurate monitoring and too high cognitive load during learning. Future research could investigate the role of support during learning to solve problems in a self-regulated way.

## ETHICS STATEMENT

In accordance with the guidelines of the ethical committee at the Department of Psychology, Education and Child studies, Erasmus University Rotterdam, the study was exempt from ethical approval procedures because the materials and procedures were not invasive.

## AUTHOR CONTRIBUTIONS

MB, LW, and FP worked together on the theoretical framework and the design of the study. Both MB and LW prepared the materials, and collected the data. MB, LW, and FP worked

together on the analysis of the data. Writing the manuscript was done by MB, LW, and FP in collaboration. Finally, MB, LW, and FP approved the manuscript and are accountable for it.

## ACKNOWLEDGMENTS

The authors would like to thank Claudin Martina for supervising the experimental sessions in the classroom. We would also like to thank the schools and teachers involved in this study for their participation.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fpsyg.2017.01346/full#supplementary-material>

## REFERENCES

- Ackerman, R., and Thompson, V. (2015). "Meta-reasoning: what can we learn from meta-memory," in *Reasoning as Memory*, eds A. Feeney and V. Thompson (Hove: Psychology Press), 164–178.
- Ariel, R., Dunlosky, J., and Bailey, H. (2009). Agenda-based regulation of study-time allocation: when agendas override item-based monitoring. *J. Exp. Psychol.* 138, 432–447. doi: 10.1037/a0015928
- Baars, M., Van Gog, T., Bruin, A., and Paas, F. (2014a). Effects of problem solving after worked examples study on primary school children's monitoring accuracy. *Appl. Cogn. Psychol.* 28, 382–391. doi: 10.1002/acp.3008
- Baars, M., Van Gog, T., De Bruin, A., and Paas, F. (2016). Effects of problem solving after worked example study on secondary school children's monitoring accuracy. *Educ. Psychol.* 37, 810–834. doi: 10.1080/01443410.2016.1150419
- Baars, M., Vink, S., Van Gog, T., De Bruin, A., and Paas, F. (2014b). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learn. Instr.* 33, 92–107. doi: 10.1016/j.learninstruc.2014.04.004
- Baars, M., Visser, S., Van Gog, T., De Bruin, A., and Paas, F. (2013). Completion of partially worked examples as a generation strategy for improving monitoring accuracy. *Contemp. Educ. Psychol.* 38, 395–406. doi: 10.1016/j.cedpsych.2013.09.001
- Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Perspective*. Englewood Cliffs, NJ: Princeton-Hall.
- Bjork, R. A., Dunlosky, J., and Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. *Annu. Rev. Psychol.* 64, 417–444. doi: 10.1146/annurev-psych-113011-143823
- Brand, S., Reimer, T., and Opwis, K. (2007). How do we learn in a negative mood? Effects of a negative mood on transfer and learning. *Learn. Instr.* 17, 1–16. doi: 10.1016/j.learninstruc.2006.11.002
- Deci, E. L., and Ryan, R. M. (2000). The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. *Psychol. Inq.* 11, 227–268. doi: 10.1207/S15327965PLI1104-01
- Dent, A. L., and Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: a meta-analysis. *Educ. Psychol. Rev.* 28, 425–474. doi: 10.1007/s10648-015-9320-8
- Diener, E., Larsen, R. J., Levine, S., and Emmons, R. A. (1985). Intensity and frequency: dimensions underlying positive and negative affect. *J. Pers. Soc. Psychol.* 48, 1253–1265. doi: 10.1037/0022-3514.48.5.1253
- Dunlosky, J., and Thiede, K. W. (2004). Causes and constraints of the shift-to-easier-materials effect in the control of study. *Mem. Cogn.* 32, 779–788. doi: 10.3758/BF03195868
- Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: the MSRL model. *Educ. Psychol.* 46, 6–25. doi: 10.1080/00461520.2011.538645
- Estrada, C. A., Isen, A. M., and Young, M. J. (1994). Positive affect improves creative problem solving and influences reported source of practice satisfaction in physicians. *Motiv. Emot.* 18, 285–299. doi: 10.1007/BF02856470
- Fiedler, K. (2001). "Affective states trigger processes," in *Theories of Mood and Cognition: A User's Guidebook*, eds L. L. Martin and G. L. Clore (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 85–98.
- Field, A. (2009). *Discovering Statistics Using SPSS*. London: Sage Publications.
- Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Think. Reason.* 7, 69–89. doi: 10.1080/13546780042000046
- García, T., Rodríguez, C., González-Castro, P., González-Pienda, J. A., and Torrance, M. (2015). Elementary students' metacognitive processes and post-performance calibration on mathematical problem-solving tasks. *Metacogn. Learn.* 11, 139–170. doi: 10.1007/s11409-015-9139-1
- Griffin, T. D., Wiley, J., and Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: concurrent processing and cue validity as constraints on metacomprehension accuracy. *Mem. Cogn.* 36, 93–103. doi: 10.3758/MC.36.1.93
- Isen, A. M. (2001). An influence of positive affect on decision making in complex situations: theoretical issues with practical implications. *J. Consum. Psychol.* 11, 75–85. doi: 10.1207/S15327663JCP1102\_01
- Jonassen, D. H. (2011). *Learning to Solve Problems: A Handbook for Designing Problem-Solving Learning Environments*. New York, NY: Routledge.
- Kalyuga, S., and Singh, A. M. (2016). Rethinking the boundaries of cognitive load theory in complex learning. *Educ. Psychol. Rev.* 28, 831–852. doi: 10.1007/s10648-015-9352-0
- Kanfer, R., and Ackerman, P. L. (1989). Motivation and cognitive abilities: an integrative/aptitude-treatment interaction approach to skill acquisition. *J. Appl. Psychol.* 74, 657–690. doi: 10.1037/0021-9010.74.4.657
- Kaufmann, G., and Vosburg, S. K. (1997). 'Paradoxical' mood effects on creative problem-solving. *Cogn. Emot.* 11, 151–170. doi: 10.1080/026999397379971
- Koriat, A., and Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *J. Exp. Psychol.* 31, 187–194. doi: 10.1037/0278-7393.31.2.187
- Koriat, A., and Bjork, R. A. (2006). Mending metacognitive illusions: a comparison of mnemonic-based and theory-based procedures. *J. Exp. Psychol.* 32, 1133–1145. doi: 10.1037/0278-7393.32.5.1133
- Koriat, A., and Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychol. Rev.* 103, 490–517. doi: 10.1037/0033-295X103.3.490



- Koriat, A., Nussinson, R., and Ackerman, R. (2014). Judgments of learning depend on how learners interpret study effort. *J. Exp. Psychol.* 40, 1624–1637. doi: 10.1037/xlm0000009
- Kornell, N., and Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *J. Exp. Psychol.* 32, 609–622. doi: 10.1037/0278-7393.32.3.609
- Kostons, D., Van Gog, T., and Paas, F. (2012). Training self-assessment and task-selection skills: a cognitive approach to improving self-regulated learning. *Learn. Instr.* 22, 121–132. doi: 10.1016/j.learninstruc.2011.08.00
- Mayer, R. E. (1992). *Thinking, Problem Solving, Cognition*. New York, NY: WH Freeman.
- Mayer, R. E. (1998). Cognitive, metacognitive, and motivational aspects of problem solving. *Instr. Sci.* 26, 49–63. doi: 10.1023/A:1003088013286
- Mega, C., Ronconi, L., and De Beni, R. (2014). What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement. *J. Educ. Psychol.* 106, 121–131. doi: 10.1037/a0033546
- Nelson, T. O., Dunlosky, J., Graf, A., and Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychol. Sci.* 5, 207–213. doi: 10.1111/j.1467-9280.1994.tb00502.x
- Osman, M. (2010). Controlling uncertainty: a review of human behavior in complex dynamic environments. *Psychol. Bull.* 136, 65–86. doi: 10.1037/a0017815
- Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *J. Educ. Psychol.* 84, 429–434. doi: 10.1037/0022-0663.84.4.429
- Paas, F., Renkl, A., and Sweller, J. (2003a). Cognitive load theory and instructional design: recent developments. *Educ. Psychol.* 38, 1–4. doi: 10.1207/S15326985EP3801-1
- Paas, F., Tuovinen, J., Tabbers, H., and Van Gerven, P. W. M. (2003b). Cognitive load measurement as a means to advance cognitive load theory. *Educ. Psychol.* 38, 63–71. doi: 10.1207/S15326985EP3801-8
- Paas, F., and Van Gog, T. (2006). Optimising worked example instruction: different ways to increase germane cognitive load. *Learn. Instr.* 16, 87–91. doi: 10.1016/j.learninstruc.2006.02.004
- Paas, F., and Van Merriënboer, J. J. (1993). The efficiency of instructional conditions: an approach to combine mental effort and performance measures. *Hum. Factors* 35, 737–743. doi: 10.1177/001872089303500412
- Pekrun, R., Goetz, T., Titz, W., and Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: a program of qualitative and quantitative research. *Educ. Psychol.* 37, 91–105. doi: 10.1207/S15326985EP3702-4
- Pintrich, P. R. (1999). The role of motivation in promoting and sustaining self-regulated learning. *Int. J. Educ. Res.* 31, 459–470. doi: 10.1016/S0883-0355(99)00015-4
- Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educ. Psychol. Rev.* 16, 385–407. doi: 10.1007/s10648-004-0006-x
- Politis, J., and Houtz, J. C. (2015). Effects of positive mood on generative and evaluative thinking in creative problem solving. *SAGE Open* 5, 1–8. doi: 10.1177/2158244015592679
- Ryan, R. M., and Deci, E. L. (2000a). Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Educ. Psychol.* 25, 54–67. doi: 10.1006/ceps.1999.1020
- Ryan, R. M., and Deci, E. L. (2000b). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* 55, 68–78. doi: 10.1037/0003-066X.55.1.68
- Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacogn. Learn.* 4, 33–45. doi: 10.1007/s11409-008-9031-3
- Schunk, D. H. (2014). *Learning Theories: An Educational Perspective*, 6th Edn. Harlow: Pearson Education Limited.
- Spering, M., Wagener, D., and Funke, J. (2005). The role of emotions in complex problem-solving. *Cogn. Emot.* 19, 1252–1261. doi: 10.1080/02699930500304886
- Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cogn. Sci.* 12, 257–285. doi: 10.1207/s15516709cog1202\_4
- Sweller, J. (2010). Element interactivity and intrinsic, extraneous and germane cognitive load. *Educ. Psychol. Rev.* 22, 123–138. doi: 10.1007/s10648-010-9128-5
- Sweller, J., Ayres, P., and Kalyuga, S. (2011). “Measuring cognitive load,” in *Cognitive Load Theory*, eds J. Sweller, P. Ayres, and S. Kalyuga (New York, NY: Springer), 71–85. doi: 10.1007/978-1-4419-8126-4\_6
- Sweller, J., and Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cogn. Instr.* 2, 59–89. doi: 10.1207/s1532690xci0201-3
- Sweller, J., and Sweller, S. (2006). Natural information processing systems. *Evol. Psychol.* 4, 434–458. doi: 10.1177/147470490600400135
- Sweller, J., Van Merriënboer, J. J. G., and Paas, F. (1998). Cognitive architecture and instructional design. *Educ. Psychol. Rev.* 10, 251–296. doi: 10.1023/A:1022193728205
- Thiede, K. W., Anderson, M., and Theriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *J. Educ. Psychol.* 95, 66–73. doi: 10.1037/0022-0663.95.1.66
- Van Gog, T., Kester, L., and Paas, F. (2011). Effects of concurrent monitoring on cognitive load and performance as a function of task complexity. *Appl. Cogn. Psychol.* 25, 584–587. doi: 10.1002/acp.1726
- Van Gog, T., Kirschner, F., Kester, L., and Paas, F. (2012). Timing and frequency of mental effort measurement: evidence in favour of repeated measures. *Appl. Cogn. Psychol.* 26, 833–839. doi: 10.1002/acp.2883
- Van Gog, T., and Paas, F. (2008). Instructional efficiency: revisiting the original construct in educational research. *Educ. Psychol.* 43, 16–26. doi: 10.1080/00461520701756248
- Van Gog, T., and Rummel, N. (2010). Example-based learning: integrating cognitive and social-cognitive research perspectives. *Educ. Psychol. Rev.* 22, 155–174. doi: 10.1007/s10648-010-9134-7
- Vansteenkiste, M., Sierens, E., Soenens, B., Luyckx, K., and Lens, W. (2009). Motivational profiles from a self-determination perspective: the quality of motivation matters. *J. Educ. Psychol.* 101, 671–688. doi: 10.1037/a0015083
- Vansteenkiste, M., Simons, J., Lens, W., Sheldon, K. M., and Deci, E. L. (2004). Motivating learning, performance, and persistence: the synergistic effects of intrinsic goal contents and autonomy-supportive contexts. *J. Pers. Soc. Psychol.* 87, 246–260. doi: 10.1037/0022-3514.87.2.246
- Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Pers. Soc. Psychol.* 54, 1063–1070. doi: 10.1037/0022-3514.54.6.1063
- Winne, P. H., and Hadwin, A. F. (1998). “Studying as self-regulated learning,” in *Metacognition in Educational Theory and Practice*, eds D. J. Hacker, J. Dunlosky, and A. C. Graesser (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 27–30.
- Zimmerman, B. J. (2008). Investigating self-regulation and motivation: historical background, methodological developments, and future prospects. *Am. Educ. Res. J.* 45, 166–183. doi: 10.3102/0002831207312909
- Zimmerman, B. J., and Campillo, M. (2003). “Motivating self-regulated problem solvers,” in *The Psychology of Problem Solving*, eds J. E. Davidson and J. Sternberg (Cambridge: Cambridge University Press), 233–262. doi: 10.1017/CBO9780511615771.009

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Baars, Wijnia and Paas. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Complex Problem Solving in Teams: The Impact of Collective Orientation on Team Process Demands

Vera Hagemann\* and Annette Kluge

Business Psychology, Faculty of Psychology, Ruhr-University Bochum, Bochum, Germany

## OPEN ACCESS

### Edited by:

Eddy J. Davelaar,  
Birkbeck University of London,  
United Kingdom

### Reviewed by:

Sorin Cristian Ionescu,  
Politehnica University of Bucharest,  
Romania  
Beno Csapo,  
University of Szeged, Hungary

### \*Correspondence:

Vera Hagemann  
vera.hagemann@rub.de

### Specialty section:

This article was submitted to  
Cognitive Science,  
a section of the journal  
Frontiers in Psychology

**Received:** 04 May 2017

**Accepted:** 19 September 2017

**Published:** 29 September 2017

### Citation:

Hagemann V and Kluge A (2017)  
Complex Problem Solving in Teams:  
The Impact of Collective Orientation  
on Team Process Demands.  
Front. Psychol. 8:1730.  
doi: 10.3389/fpsyg.2017.01730

Complex problem solving is challenging and a high-level cognitive process for individuals. When analyzing complex problem solving in teams, an additional, new dimension has to be considered, as teamwork processes increase the requirements already put on individual team members. After introducing an idealized teamwork process model, that complex problem solving teams pass through, and integrating the relevant teamwork skills for interdependently working teams into the model and combining it with the four kinds of team processes (transition, action, interpersonal, and learning processes), the paper demonstrates the importance of fulfilling team process demands for successful complex problem solving within teams. Therefore, results from a controlled team study within complex situations are presented. The study focused on factors that influence action processes, like coordination, such as emergent states like collective orientation, cohesion, and trust and that dynamically enable effective teamwork in complex situations. Before conducting the experiments, participants were divided by median split into two-person teams with either high ( $n = 58$ ) or low ( $n = 58$ ) collective orientation values. The study was conducted with the microworld C<sup>3</sup>Fire, simulating dynamic decision making, and acting in complex situations within a teamwork context. The microworld includes interdependent tasks such as extinguishing forest fires or protecting houses. Two firefighting scenarios had been developed, which takes a maximum of 15 min each. All teams worked on these two scenarios. Coordination within the team and the resulting team performance were calculated based on a log-file analysis. The results show that no relationships between trust and action processes and team performance exist. Likewise, no relationships were found for cohesion. Only collective orientation of team members positively influences team performance in complex environments mediated by action processes such as coordination within the team. The results are discussed in relation to previous empirical findings and to learning processes within the team with a focus on feedback strategies.

**Keywords:** interdependence, team processes, complex problem solving, collective orientation, trust, cohesion, C<sup>3</sup>Fire, microworld

## INTRODUCTION

Complex problems in organizational contexts are seldom solved by individuals. Generally, interdependently working teams of experts deal with complex problems (Fiore et al., 2010), which are characterized by element interactivity/ interconnectedness, dynamic developments, non-transparency and multiple, and/or conflicting goals (Dörner et al., 1983; Brehmer, 1992; Funke, 1995). Complex problem solving “takes place for reducing the barrier between a given start state and an intended goal state with the help of cognitive activities and behavior. Start state, intended goal state, and barriers prove complexity, change dynamically over time, and can be partially intransparent” (Funke, 2012, p. 682). Teams dealing with complex problems in interdependent work contexts, for example in disaster, crisis or accident management, are called High Responsibility Teams. They are named High Responsibility Teams (HRTs; Hagemann, 2011; Hagemann et al., 2011) due to their dynamic and often unpredictable working conditions and demanding work contexts, in which technical faults and slips have severe consequences for human beings and the environment if they are not identified and resolved within the team immediately (Kluge et al., 2009). HRTs bear responsibility regarding lives of third parties and their own lives based on their actions and consequences.

The context of interdependently working HRTs, dealing with complex problems, is described as follows (Zsombok, 1997): Members of interdependently working teams have to reach ill-defined or competing goals in common in poor structured, non-transparent and dynamically changing situations under the consideration of rules of engagement and based on several cycles of joint action. Some or all goals are critical in terms of time and the consequences of actions result in decision-based outcomes with high importance for the culture (e.g., human life). In HRT contexts, added to the features of the complexity of the problem, is the complexity of relationships, which is called social complexity (Dörner, 1989/2003) or crew coordination complexity (Kluge, 2014), which results from the interconnectedness between multiple agents through coordination requirements. The dynamic control aspect of the continuous process is coupled with the need to coordinate multiple highly interactive processes imposing high coordination demands (Roth and Woods, 1988; Waller et al., 2004; Hagemann et al., 2012).

Within this article, it is important to us to describe the theoretical background of complex problem solving in teams in depth and to combine different but compatible theoretical approaches, in order to demonstrate their theoretical and practical use in the context of the analysis of complex problem solving in teams. In Industrial and Organizational Psychology, a detailed description of tasks and work contexts that are in the focus of the analysis is essential. The individual or team task is the point of intersection between organization and individual as a “psychologically most relevant part” of the working conditions (Ulich, 1995). Thus, the tasks and the teamwork context of teams that deal with complex problems is of high relevance in the present paper. We will comprehensively describe the context of

complex problem solving in teams by introducing a model of an idealized teamwork process that complex problem solving teams pass through and extensively integrate the relevant teamwork skills for these interdependently working teams into the idealized teamwork process model.

Furthermore, we will highlight the episodic aspect concerning complex problem solving in teams and combine the agreed on transition, action, interpersonal and learning processes of teamwork with the idealized teamwork process model. Because we are interested in investigating teamwork competencies and action processes of complex problem solving teams, we will analyze the indirect effect of collective orientation on team performance through the teams’ coordination behavior. The focusing of the study will be owed to its validity. Even though that we know that more aspects of the theoretical framework might be of interest and could be analyzed, we will focus on a detail within the laboratory experiment for getting reliable and valid results.

## Goal, Task, and Outcome Interdependence in Teamwork

Concerning interdependence, teamwork research focuses on three designated features, which are in accordance with general process models of human action (Hertel et al., 2004). One type is *goal* interdependence, which refers to the degree to which teams have distinct goals as well as a linkage between individual members and team goals (Campion et al., 1993; Wageman, 1995). A second type is *task* interdependence, which refers to the interaction between team members. The team members depend on each other for work accomplishment, and the actions of one member have strong implications for the work process of all members (Shea and Guzzo, 1987; Campion et al., 1993; Hertel et al., 2004). The third type is *outcome* interdependence, which is defined as the extent to which one team member’s outcomes depend on the performance of other members (Wageman, 1995). Accordingly, the rewards for each member are based on the total team performance (Hertel et al., 2004). This can occur, for instance, if a team receives a reward based on specific performance criteria. Although interdependence is often the reason why teams are formed in the first place, and it is stated as a defining attribute of teams (Salas et al., 2008), different levels of task interdependence exist (Van de Ven et al., 1976; Arthur et al., 2005).

The workflow pattern of teams can be

- (1) Independent or pooled (activities are performed separately),
- (2) Sequential (activities flow from one member to another in a unidirectional manner),
- (3) Reciprocal (activities flow between team members in a back and forth manner) or
- (4) Intensive (team members must simultaneously diagnose, problem-solve, and coordinate as a team to accomplish a task).

Teams that deal with complex problems work within intensive interdependence, which requires greater coordination patterns compared to lower levels of interdependence (Van de Ven et al., 1976; Wageman, 1995) and necessitates mutual adjustments as

well as frequent interaction and information integration within the team (Gibson, 1999; Stajkovic et al., 2009).

Thus, in addition to the cognitive requirements related to information processing (e.g., encoding, storage and retrieval processes (Hinsz et al., 1997), simultaneously representing and anticipating the dynamic elements and predicting future states of the problem, balancing contradictory objectives and decide on the right timing for actions to execute) of individual team members, the interconnectedness between the experts in the team imposes high team process demands on the team members. These team process demands follow from the required interdependent actions of all team members for effectively using all resources, such as equipment, money, time, and expertise, to reach high team performance (Marks et al., 2001). Examples for team process demands are the communication for building a shared situation awareness, negotiating conflicting perspectives on how to proceed or coordinating and orchestrating actions of all team members.

## A Comprehensive Model of the Idealized Teamwork Process

The cognitive requirements, that complex problem solving teams face, and the team process demands are consolidated within our model of an idealized teamwork process in **Figure 1** (Hagemann, 2011; Kluge et al., 2014). Individual and team processes converge sequential and in parallel and influencing factors as well as process demands concerning complex problem solving in teams can be extracted. The core elements of the model are situation awareness, information transfer, individual and shared mental models, coordination and leadership, and decision making.

Complex problem solving teams are responsible for finding solutions and reaching specified goals. Based on the overall goals various sub goals will be identified at the beginning of the teamwork process in the course of mission analysis, strategy formulation and planning, all aspects of the *transition phase* (Marks et al., 2001). The transition phase processes occur during periods of time when teams focus predominantly on evaluation and/or planning activities. The identified and communicated goals within the team represent relevant input variables for each team member in order to build up a *Situation Awareness* (SA). SA contains three steps and is the foundation for an ideal and goal directed collaboration within a team (Endsley, 1999; Flin et al., 2008). The individual SA is the start and end within the idealized teamwork process model. SA means the assessment of a situation which is important for complex problem solving teams, as they work based on the division of labor as well as interdependently and each team member needs to achieve a correct SA and to share it within the team. Each single team member needs to utilize all technical and interpersonal resources in order to collect and interpret up-to-date goal directed information and to share this information with other team members via “closed-loop communication.”

This *information transfer* focuses on sending and receiving single SA between team members in order to build up a *Shared Situation Awareness* (SSA). Overlapping cuts of individual SA are synchronized within the team and a bigger picture of the

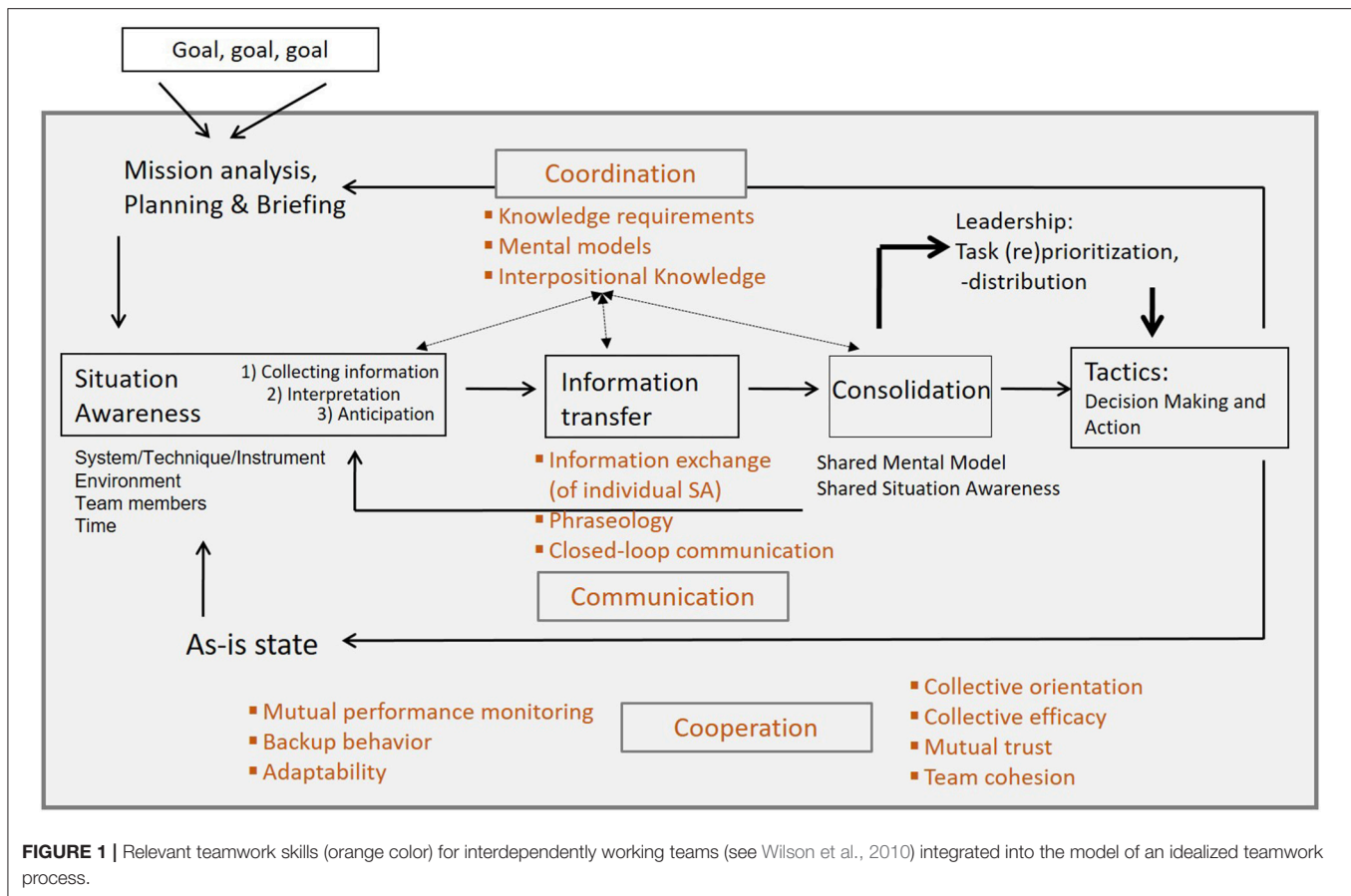
situation is developed. Creating a SSA means sharing a common perspective of the members concerning current events within their environment, their meaning and their future development. This shared perspective enables problem-solving teams to attain high performance standards through corresponding and goal directed actions (Cannon-Bowers et al., 1993).

Expectations of each team member based on briefings, *individual mental models* and *interpositional knowledge* influence the SA, the information transfer and the consolidation process. Mental models are internal and cognitive representations of relations and processes (e.g., execution of tactics) between various aspects or elements of a situation. They help team members to describe, explain and predict circumstances (Mathieu et al., 2000). Mental models possess knowledge elements required by team members in order to assess a current situation in terms of SA. Interpositional knowledge refers to an individual understanding concerning the tasks and duties of all team members, in order to develop an understanding about the impact of own actions on the actions of other team members and vice versa. It supports the team in identifying the information needs and the amount of required help of other members and in avoiding team conflicts (Smith-Jentsch et al., 2001). This knowledge is the foundation for anticipating the team members’ needs for information and it is important for matching information within the team.

Based on the information matching process within the team, a common understanding of the problem, the goals and the current situation is developed in terms of a *Shared Mental Model* (SMM), which is important for the subsequent decisions. SMM are commonly shared mental models within a team and refer to the organized knowledge structures of all team members, that are shared with each other and which enable the team to interact goal-oriented (Mathieu et al., 2000). SMM help complex problem solving teams during high workload to adapt fast and efficiently to changing situations (Waller et al., 2004). They also enhance the teams’ performance and communication processes (Cannon-Bowers et al., 1993; Mathieu et al., 2000). Especially under time pressure and in crucial situations when overt verbal communication and explicit coordination is not applicable, SMM are fundamental in order to coordinate implicitly. This information matching process fosters the building of a shared understanding of the current situation and the required actions. In order to do so teamwork skills (see Wilson et al., 2010) such as *communication*, *coordination*, and *cooperation* within the team are vitally important. **Figure 1** incorporates the teamwork skills into the model of an idealized teamwork process.

Depending on the shared knowledge and SA within the team, the coordination can be based either on well-known procedures or shared expectations within the team or on explicit communication based on task specific phraseology or closed-loop communication. Cooperation needs mutual performance monitoring within the team, for example, in order to apply task strategies to accurately monitor teammate performance and prevent errors (Salas et al., 2005). Cooperation also needs backup behavior of each team member, for example, and continuous actions in reference to the collective events. The anticipation of other team members’ needs under high workload maintains the





teams' performance and the well-being of each team member (Badke-Schaub, 2008). A successful pass through the teamwork process model also depends e.g., on the trust and the cohesion within the team and the collective orientation of each team member.

Collective orientation (CO) is defined "as the propensity to work in a collective manner in team settings" (Driskell et al., 2010, p. 317). Highly collectively oriented people work with others on a task-activity and team-activity track (Morgan et al., 1993) in a goal-oriented manner, seek others' input, contribute to team outcomes, enjoy team membership, and value cooperativeness more than power (Driskell et al., 2010). Thus, teams with collectively oriented members perform better than teams with non-collectively oriented members (Driskell and Salas, 1992). CO, trust and cohesion as well as other coordination and cooperation skills are so called emergent states that represent cognitive, affective, and motivational states, and not traits, of teams and team members, and which are influenced, for example, by team experience, so that emergent states can be considered as team inputs but also as team outcomes (Marks et al., 2001).

Based on the information matching process the complex problem solving team or the team leader needs to make decisions in order to execute actions. The *task prioritization* and *distribution* is an integrated part of this step (Waller et al., 2004). Depending on the progress of the dynamic, non-transparent

and heavily foreseeable situation tasks have to be re-prioritized during episodes of teamwork. Episodes are "temporal cycles of goal-directed activity" in which teams perform (Marks et al., 2001, p. 359). Thus, the team acts adaptive and is able to react flexible to situation changes. The team coordinates implicitly when each team member knows what he/she has to do in his/her job, what the others expect from him/her and how he/she interacts with the others. In contrast, when abnormal events occur and they are recognized during SA processes, the team starts coordinating explicitly via communication, for example. Via closed-loop communication and based on interpositional knowledge new strategies are communicated within the team and tasks are re-prioritized.

The result of the decision making and action taking flows back into the individual SA and the as-is state will be compared with the original goals. This model of an idealized teamwork process (Figure 1) is a regulator circuit with feedback loops, which enables a team to adapt flexible to changing environments and goals. The foundation of this model is the classic Input-Process-Outcome (IPO) framework (Hackman, 1987) with a strong focus on the process part. IPO models view processes as mechanisms linking variables such as member, team, or organizational features with outcomes such as performance quality and quantity or members' reactions. This mediating mechanism, the *team process*, can be defined as "members' interdependent acts that

convert inputs to outcomes through cognitive, verbal, and behavioral activities directed toward organizing taskwork to achieve collective goals” (Marks et al., 2001, p. 357). That means team members interact interdependently with other members as well as with their environment. These cognitive, verbal, and behavioral activities directed toward taskwork and goal attainment are represented as gathering situation awareness, communication, coordination, cooperation, the consolidation of information, and task prioritization within our model of an idealized teamwork process. Within the context of complex problem solving, teams have to face team process demands in addition to cognitive challenges related to individual information processing. That means teamwork processes and taskwork to solve complex problems co-occur, the processes guide the execution of taskwork.

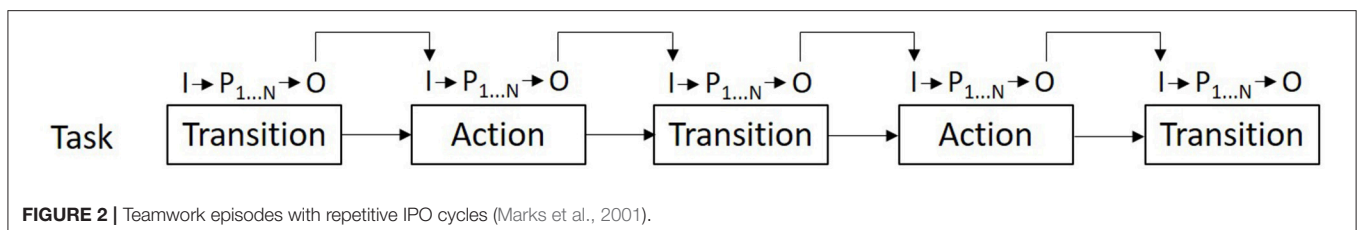
The dynamic nature of teamwork and temporal influences on complex problem solving teams are considered within adapted versions (Marks et al., 2001; Ilgen et al., 2005) of the original IPO framework. These adaptations propose that teams experience cycles of joint action, so called episodes, in which teams perform and also receive feedback for further actions. The IPO cycles occur sequentially and simultaneously and are nested in transition and action phases within episodes in which outcomes from initial episodes serve as inputs for the next cycle (see **Figure 2**). These repetitive IPO cycles are a vital element of our idealized teamwork process model, as it incorporates feedback loops in such a way, that the outcomes, e.g., changes within the as-is state, are continuously compared with the original goals. Detected discrepancies within the step of updating SA motivate the team members to consider further actions for goal accomplishment.

When applying this episodic framework to complex problem solving teams it becomes obvious that teams handle different types of taskwork at different phases of task accomplishment (Marks et al., 2001). That means episodes consist of two phases, so-called *action* and *transition* phases, in which teams are engaged in activities related to goal attainment and in other time in reflecting on past performance and planning for further common actions. The addition of the social complexity to the complexity of the problem within collaborative complex problem solving comes to the fore here. During transition phases teams evaluate their performance, compare the as-is state against goals, reflect on their strategies and plan future activities to guide their goal accomplishment. For example, team members discuss alternative courses of action, if their activities for simulated firefighting, such as splitting team members in order to cover more space of the map, are not successful. During action phases, teams focus

directly on the taskwork and are engaged in activities such as exchanging information about the development of the dynamic situation or supporting each other. For example, a team member recognizes high workload of another team member and supports him/her in collecting information or in taking over the required communication with other involved parties.

## Transition and Action Phases

The idealized teamwork process model covers these transition and action phases as well as the processes occurring during these two phases of team functioning, which can be clustered into transition, action, and interpersonal processes. That means during complex problem solving the relevant or activated teamwork processes in the transition and action phases change as teams move back and forth between these phases. As this taxonomy of team processes from Marks et al. (2001) states that a team process is multidimensional and teams use different processes simultaneously, some processes can occur either during transition periods or during action periods or during both periods. *Transition processes* especially occur during transition phases and enable the team to understand their tasks, guide their attention, specify goals and develop courses of action for task accomplishment. Thus, transition processes include (see Marks et al., 2001) mission analysis, formulation and planning (Prince and Salas, 1993), e.g., fighting a forest fire, goal specification (Prussia and Kinicki, 1996), e.g., saving as much houses and vegetation as possible, and strategy formulation (Prince and Salas, 1993; Cannon-Bowers et al., 1995), e.g., spreading team members into different geographic directions. *Action processes* predominantly occur during action phases and support the team in conducting activities directly related to goal accomplishment. Thus, action processes are monitoring progress toward goals (Cannon-Bowers et al., 1995), e.g., collecting information how many cells in a firefighting simulation are still burning, systems monitoring (Fleishman and Zaccaro, 1992), e.g., tracking team resources such as water for firefighting, team monitoring and backup behavior (Stevens and Campion, 1994; Salas et al., 2005), e.g., helping a team member and completing a task for him/her, and coordination (Fleishman and Zaccaro, 1992; Serfaty et al., 1998), e.g., orchestrating the interdependent actions of the team members such as exchanging information during firefighting about positions of team members for meeting at the right time at the right place in order to refill the firefighters water tanks. Especially the coordination process is influenced by the amount of task interdependence as coordination becomes more and more important for effective team functioning when interdependence increases (Marks et al., 2001). *Interpersonal*



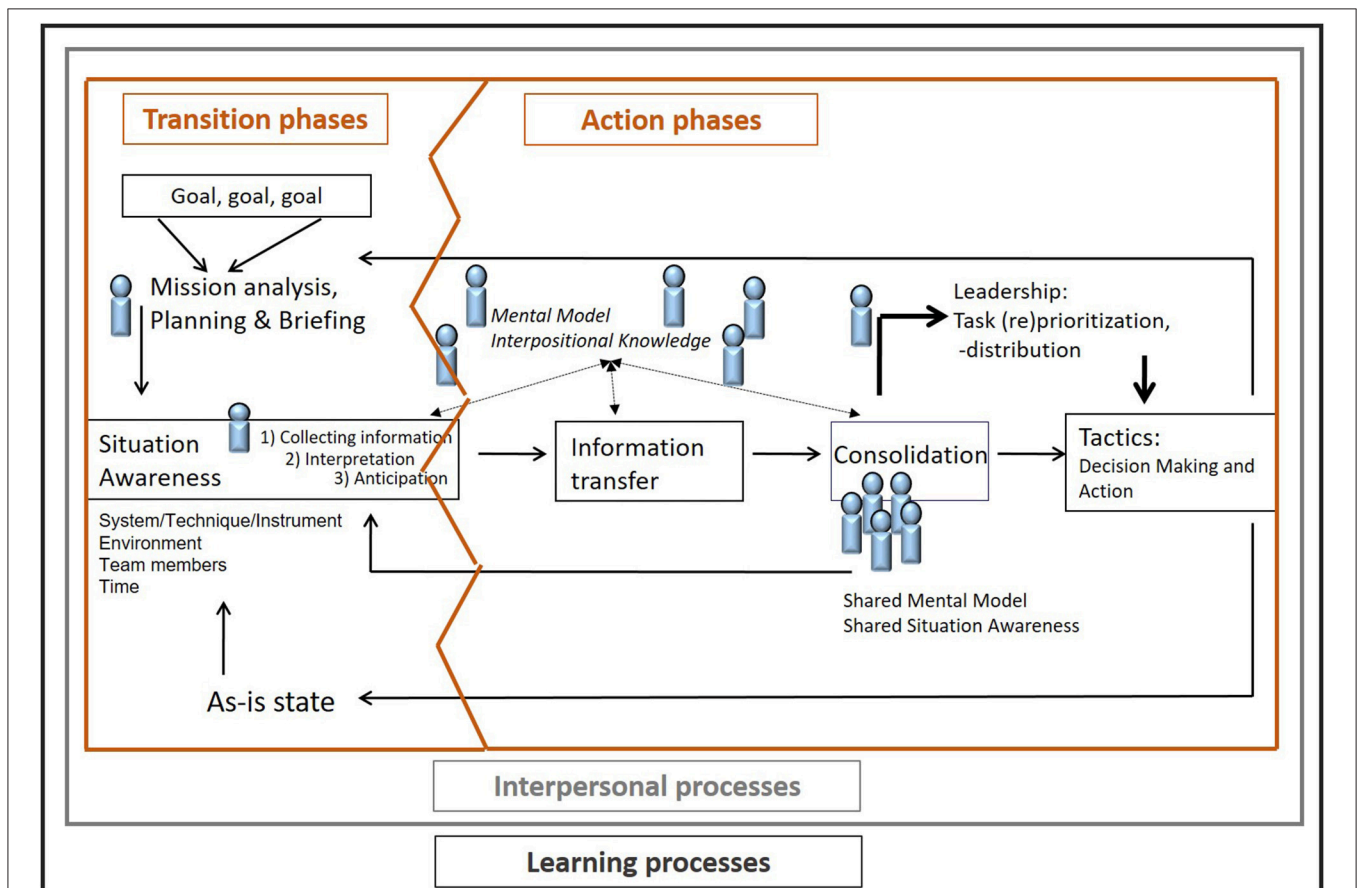
processes occur during transition and action phases equally and lay the foundation for the effectiveness of other processes and govern interpersonal activities (Marks et al., 2001). Thus, interpersonal processes include conflict management (Cannon-Bowers et al., 1995), like the development of team rules, motivation and confidence building (Fleishman and Zaccaro, 1992), like encourage team members to perform better, and affect management (Cannon-Bowers et al., 1995), e.g., regulating member emotions during complex problem solving.

Summing up, process demands such as transition processes that complex problem solving teams pass through, are mission analysis, planning, briefing and goal specification, visualized on the left side of the idealized teamwork process model (see Figure 3). The results of these IPO cycles lay the foundation for gathering a good SA and initiating activities directed toward taskwork and goal accomplishment and therefore initiating action processes. The effective execution of action processes depends on the communication, coordination, cooperation, matching of information, and task prioritization as well as emergent team cognition variables (SSA and SMM) within the team. The results, like decisions, of these IPO cycles flow back into the next episode and may initiate further transition processes. In addition, interpersonal processes play a crucial

role for complex problem solving teams. That means, conflict management, motivating and confidence building, and affect management are permanently important, no matter whether a team runs through transition or action phases and these interpersonal processes frame the whole idealized teamwork process model. Therefore, interpersonal processes are also able to impede successful teamwork at any point as breakdowns in conflict or affect management can lead to coordination breakdowns (Wilson et al., 2010) or problems with monitoring or backing up teammates (Marks et al., 2001). Thus, complex problem solving teams have to face these multidimensional team process demands in addition to cognitive challenges, e.g., information storage or retrieval (Hinsz et al., 1997), related to individual information processing.

### Team Learning Opportunities for Handling Complex Problems

In order to support teams in handling complex situations or problems, learning opportunities seem to be very important for successful task accomplishment and for reducing possible negative effects of team process demands. Learning means any kind of relative outlasted changes in potential of human behavior that cannot be traced back to age-related changes (Bower and



**FIGURE 3 |** The integration of transition, action, interpersonal, and learning processes into the model of an idealized teamwork process.

Hilgard, 1981; Bredenkamp, 1998). Therefore, Schmutz et al. (2016) amended the taxonomy of team processes developed by Marks et al. (2001) and added learning processes as a fourth category of processes, which occur during transition and action phases and contribute to overall team effectiveness. *Learning processes* (see also Edmondson, 1999) include observation, e.g., observing own and other team members' actions such as the teammate's positioning of firewalls in order to protect houses in case of firefighting, feedback, like giving a teammate information about the wind direction for effective positioning of firewalls, and reflection, e.g., talking about procedures for firefighting or refilling water tanks, for example, within the team. Learning from success and failure and identifying future problems is crucial for the effectiveness of complex problem solving teams and therefore possibilities for learning based on repetitive cycles of joint action or episodes and reflection of team members' activities during action and transition phases should be used effectively (Edmondson, 1999; Marks et al., 2001). The processes of the idealized teamwork model are embedded into these learning processes (see **Figure 3**).

The fulfillment of transition, action, interpersonal and learning processes contribute significantly to successful team performance in complex problem solving. For clustering these processes, transition and action processes could be seen as operational processes and interpersonal and learning process as support processes. When dealing with complex and dynamic situations teams have to face these team process demands more strongly than in non-complex situations. For example, goal specification and prioritization or strategy formulation, both aspects of transition processes, are strongly influenced by multiple goals, interconnectedness or dynamically and constantly changing conditions. The same is true for action processes, such as monitoring progress toward goals, team monitoring and backup behavior or coordination of interdependent actions. Interpersonal processes, such as conflict and affect management or confidence building enhance the demands put on team members compared to individuals working on complex problems. Interpersonal processes are essential for effective teamwork and need to be cultivated during episodes of team working, because breakdowns in confidence building or affect management can lead to coordination breakdowns or problems with monitoring or backing up teammates (Marks et al., 2001). Especially within complex situations aspects such as interdependence, delayed feedback, multiple goals and dynamic changes put high demands on interpersonal processes within teams. Learning processes, supporting interpersonal processes and the result of effective teamwork are e.g., observation of others' as well as own actions and receiving feedback by others or the system and are strongly influenced by situational characteristics such as non-transparency or delayed feedback concerning actions. It is assumed that amongst others team learning happens through repetitive cycles of joint action within the action phases and reflection of team members within the transition phases (Edmondson, 1999; Gabelica et al., 2014; Schmutz et al., 2016). The repetitive cycles help to generate SMM (Cannon-Bowers et al., 1993; Mathieu et al., 2000), SSA

(Endsley and Robertson, 2000) or transactive memory systems (Hollingshead et al., 2012) within the team.

## Emergent States in Complex Team Work and the Role of Collective Orientation

IPO models propose that *input variables* and *emergent states* are able to influence team processes and therefore outcomes such as team performance positively. Emergent states represent team members' attitudes or motivations and are "properties of the team that are typically dynamic in nature and vary as a function of team context, inputs, processes, and outcomes" (Marks et al., 2001, p. 357). Both emergent states and interaction processes are relevant for team effectiveness (Kozlowski and Ilgen, 2006).

Emergent states refer to conditions that underlie and dynamically enable effective teamwork (DeChurch and Mesmer-Magnus, 2010) and can be differentiated from team process, which refers to interdependent actions of team members that transform inputs into outcomes based on activities directed toward task accomplishment (Marks et al., 2001). Emergent states mainly support the execution of behavioral processes (e.g., planning, coordination, backup behavior) during the action phase, meaning during episodes when members are engaged in acts that focus on task work and goal accomplishment. Emergent states like trust, cohesion and CO are "products of team experiences (including team processes) and become new inputs to subsequent processes and outcomes" (Marks et al., 2001, p. 358). Trust between team members and cohesion within the team are emergent states that develop over time and only while experiencing teamwork in a specific team. CO is an emergent state that a team member brings along with him/her into the teamwork, is assumed to be more persistent than trust and cohesion, and can, but does not have to, be positively and negatively influenced by experiencing teamwork in a specific team for a while or by means of training (Eby and Dobbins, 1997; Driskell et al., 2010). Thus, viewing emergent states on a continuum, trust and cohesion are assumed more fluctuating than CO, but CO is much more sensitive to change and direct experience than a stable trait such as a personality trait.

CO of team members is one of the teamwork-relevant competencies that facilitates team processes, such as collecting and sharing information between team members, and positively affects the success of teams, as people who are high in CO work with others in a goal-oriented manner, seek others' input and contribute to team outcomes (Driskell et al., 2010). CO is an emergent state, as it can be an input variable as well as a teamwork outcome. CO is context-dependent, becomes visible in reactions to situations and people, and can be influenced by experience (e.g., individual learning experiences with various types of teamwork) or knowledge or training (Eby and Dobbins, 1997; Bell, 2007). CO enhances team performance through activating transition and action processes such as coordination, evaluation and consideration of task inputs from other team members while performing a team task (Driskell and Salas, 1992; Salas et al., 2005). Collectively oriented people effectively use available resources in due consideration of the team's goals,



participate actively and adapt teamwork processes adequately to the situation.

Driskell et al. (2010) and Hagemann (2017) provide a sound overview of the evidence of discriminant and convergent validity of CO compared to other teamwork-relevant constructs, such as cohesion, also an emergent state, or cooperative interdependence or preference for solitude. Studies analyzing collectively and non-collectively oriented persons' decision-making in an interdependent task demonstrated that teams with non-collectively oriented members performed poorly in problem solving and that members with CO judged inputs from teammates as more valuable and considered these inputs more frequently (Driskell and Salas, 1992). Eby and Dobbins (1997) also showed that CO results in increased coordination among team members, which may enhance team performance through information sharing, goal setting and strategizing (Salas et al., 2005). Driskell et al. (2010) and Hagemann (2017) analyzed CO in relation to team performance and showed that the effect of CO on team performance depends on the task type (see McGrath, 1984). Significant positive relationships between team members' CO and performance were found in relation to the task types choosing/decision making and negotiating (Driskell et al., 2010) respectively choosing/decision making (Hagemann, 2017). These kinds of tasks are characterized by much more interdependence than task types such as executing or generating tasks. As research shows that the positive influence of CO on team performance unfolds especially in interdependent teamwork contexts (Driskell et al., 2010), which require more team processes such as coordination patterns (Van de Ven et al., 1976; Wageman, 1995) and necessitate mutual adjustments as well as frequent information integration within the team (Gibson, 1999; Stajkovic et al., 2009), CO might be vitally important for complex problem solving teams. Thus, CO as an emergent state of single team members might be a valuable resource for enhancing the team's performance when exposed to solving complex problems. Therefore, it will be of interest to analyze the influence of CO on team process demands such as coordination processes and performance within complex problem solving teams. We predict that the positive effect of CO on team performance is an indirect effect through coordination processes within the team, which are vitally important for teams working in intensive interdependent work contexts.

Hypothesis 1: CO leads to a better coordination behavior, which in turn leads to a higher team performance.

As has been shown in team research that emergent states like trust and cohesion (see also **Figure 1**) affect team performance, these two constructs are analyzed in conjunction with CO concerning action processes, such as coordination behavior and team performance. Trust between team members supports information sharing and the willingness to accept feedback, and therefore positively influences teamwork processes (McAllister, 1995; Salas et al., 2005). Cohesion within a team facilitates motivational factors and group processes like coordination and enhances team performance (Beal et al., 2003; Kozlowski and Ilgen, 2006).

Hypothesis 2: Trust shows a positive relationship with (a) action processes (team coordination) and with (b) team performance.

Hypothesis 3: Cohesion shows a positive relationship with (a) action processes (team coordination) and with (b) team performance.

## MATERIALS AND METHODS

In order to demonstrate the importance of team process demands for complex problem solving in teams, we used a computer-based microworld in a laboratory study. We analyzed the effectiveness of complex problem solving teams while considering the influence of input variables, like collective orientation of team members and trust and cohesion within the team, on action processes within teams, like coordination.

### The Microworld for Investigating Teams Process Demands

We used the simulation-based team task C<sup>3</sup>Fire (Granlund et al., 2001; Granlund and Johansson, 2004), which is described as an intensive interdependence team task for complex problem solving (Arthur et al., 2005). C<sup>3</sup>Fire is a command, control and communications simulation environment that allows teams' coordination and communication in complex and dynamic environments to be analyzed. C<sup>3</sup>Fire is a microworld, as important characteristics of the real world are transferred to a small and well-controlled simulation system. The task environment in C<sup>3</sup>Fire is complex, dynamic and opaque (see **Table 1**) and therefore similar to the cognitive tasks people usually encounter in real-life settings, in and outside their work place (Brehmer and Dörner, 1993; Funke, 2001). **Figure 4** demonstrates how the complexity characteristics mentioned in **Table 1** are realized in C<sup>3</sup>Fire. The screenshot represents the simulation manager's point of view, who is able to observe all units and actions and the scenario development. For more information about the units and scenarios, please (see the text below and the Supplementary Material). Complexity requires people to consider a number of facts. Because executed actions in C<sup>3</sup>Fire influence the ongoing process, the sequencing of actions is free and not stringent, such as a fixed (if X then Y) or parallel (if X then Y and Z) sequence (Ormerod et al., 1998). This can lead to stressful situations. Taking these characteristics of microworlds into consideration, team processes during complex problem solving can be analyzed within laboratories under controlled conditions. Simulated microworlds such as C<sup>3</sup>Fire allow the gap to be bridged between laboratory studies, which might show deficiencies regarding ecological validity, and field studies, which have been criticized due to their small amount of control (see Brehmer and Dörner, 1993).

In C<sup>3</sup>Fire, the teams' task is to coordinate their actions to extinguish a forest fire whilst protecting houses and saving lives. The team members' actions are interdependent. The simulation includes, e.g., forest fires, houses, tents, gas tanks, different

**TABLE 1** | Overview of complexity characteristics of microworlds in general and in C<sup>3</sup>Fire (cf. Funke, 2001).

Complexity	General	C <sup>3</sup> Fire examples	Representation in Figure 4
Goals	People try to reach many goals, some of which may be contradictory, and therefore they have to make trade-offs.	Extinguish a forest fire and/or protect houses simultaneously.	Two fires are spreading out. Brown cells are extinguished, black cells are burned down. A house and a school are blocked with fire-breaks (gray cells).
Side-effects	Side effects of a given course of action exist due to coupled processes and force people to choose between many possible courses of action.	If the participant decides to refill his/her water tank on his/her back, he/she is not able to fight a fire during this refill process.	Unit 2, one firefighting unit, stands on the local water tank for refilling its water supply.
Dynamic	Microworlds are dynamic, because “their current state is a function of the history of the interaction between the subject and the system” and “they change, both as a consequence of the subject’s actions and autonomously” (Brehmer and Dörner, 1993, p. 173). People have to act in real time and directly influence the system’s state even though they do not know exactly when they have to make decisions.	If the participant does nothing, the fire spreads in all directions. If the participant extinguishes burning fields, the fire spreads in the directions where no firefighting occurs. If the wind direction changes, the direction of fire spreading also changes and the participant needs to recognize this for his/her further actions.	Two fires are spreading out into all directions. The fire stops before a placed fire-break. The fire spreads out predominantly in a westward direction, because the wind is coming from the East.
Opaque	Opaque means that the people do not have all relevant information. Thus, people have to form hypotheses and test them autonomously during activity.	Restricted visibility field. Not everything within the simulation environment is visible for the participants without exploring the environment. All units see the houses, trees, bushes and so on, but they can only see the fire if they are close to it.	The restricted visibility field is represented by the yellow squares. e.g., unit 5 only sees five burning cells and four non-burning cells and has an intersection of two cells with unit 4. Unit 1 only sees eight burning cells and one burned-out cell and has an intersection of one cell with unit 4.

kinds of vegetation and computer-simulated agents such as firefighting units (Granlund, 2003). It is possible, for example, that the direction of wind will change during firefighting and the time until different kinds of vegetation are burned down varies between those. In the present study, two simulation scenarios were developed for two-person teams and consisted of two firefighting units, one mobile water tank unit (responsible for re-filling the firefighting units’ water tanks that contain a predefined amount of water) and one fire-break unit (a field defended with a fire-break cannot be ignited; the fire spreads around its ends). The two developed scenarios lasted for 15 min maximum. Each team member was responsible for two units in each scenario; person one for firefighting and water tank unit and person two for firefighting and fire-break unit. The user interface was a map system (40 × 40 square grid) with all relevant geographic information and positions of all symbols representing houses, water tank units and so on. All parts of the map with houses and vegetation were visible for the subjects, but not the fire itself or the other units; instead, the subjects were close to them with their own units (restricted visibility field; 3 × 3 square grid). The simulation was run on computers networked in a client-server configuration. The subjects used a chat system for communication that was logged. For each scenario, C<sup>3</sup>Fire creates a detailed log file containing all events that occurred over the course of the simulation. Examples of the C<sup>3</sup>Fire scenarios are provided in the Figures S1–3 and a short introduction into the microworld is given in the video. Detailed information regarding the scenario characteristics are given in Table S1. From scenario one to two, the complexity and interdependence increased.

## Participants

The study was conducted from Mai 2014 until March 2015. Undergraduate and graduate students ( $N = 116$ ) studying applied cognitive sciences participated in the study (68.1% female). Their mean age was 21.17 years ( $SD = 3.11$ ). Participants were assigned to 58 two-person teams, with team assignments being based on the pre-measured CO values (see procedure). They received 2 hourly credits as a trial subject and giveaways such as pencils and non-alcoholic canned drinks. The study was approved by the university’s ethics committee in February 2014.

## Procedure

The study was conducted within a laboratory setting at a university department for business psychology. Prior to the experiment, the participants filled in the CO instrument online and gave written informed consent (see Figure 5). The median was calculated subsequently ( $Md = 3.12$ ; range: 1.69–4.06; scale range: 1–5) relating to the variable CO and two individuals with either high ( $n = 58$ ) or low ( $n = 58$ ) CO values were randomly matched as teammates. The matching process was random in part, as those two subjects were matched to form a team, whose preferred indicated time for participation in a specific week during data collection were identical. The participants were invited to the experimental study by e-mail 1–2 weeks after filling in the CO instrument. The study began with an introduction to the experimental procedure and the teams’ task. The individuals received time to familiarize themselves with the simulation, received 20 min of training and completed two practice trials. After the training, participants answered a questionnaire collecting demographic data. Following this, a

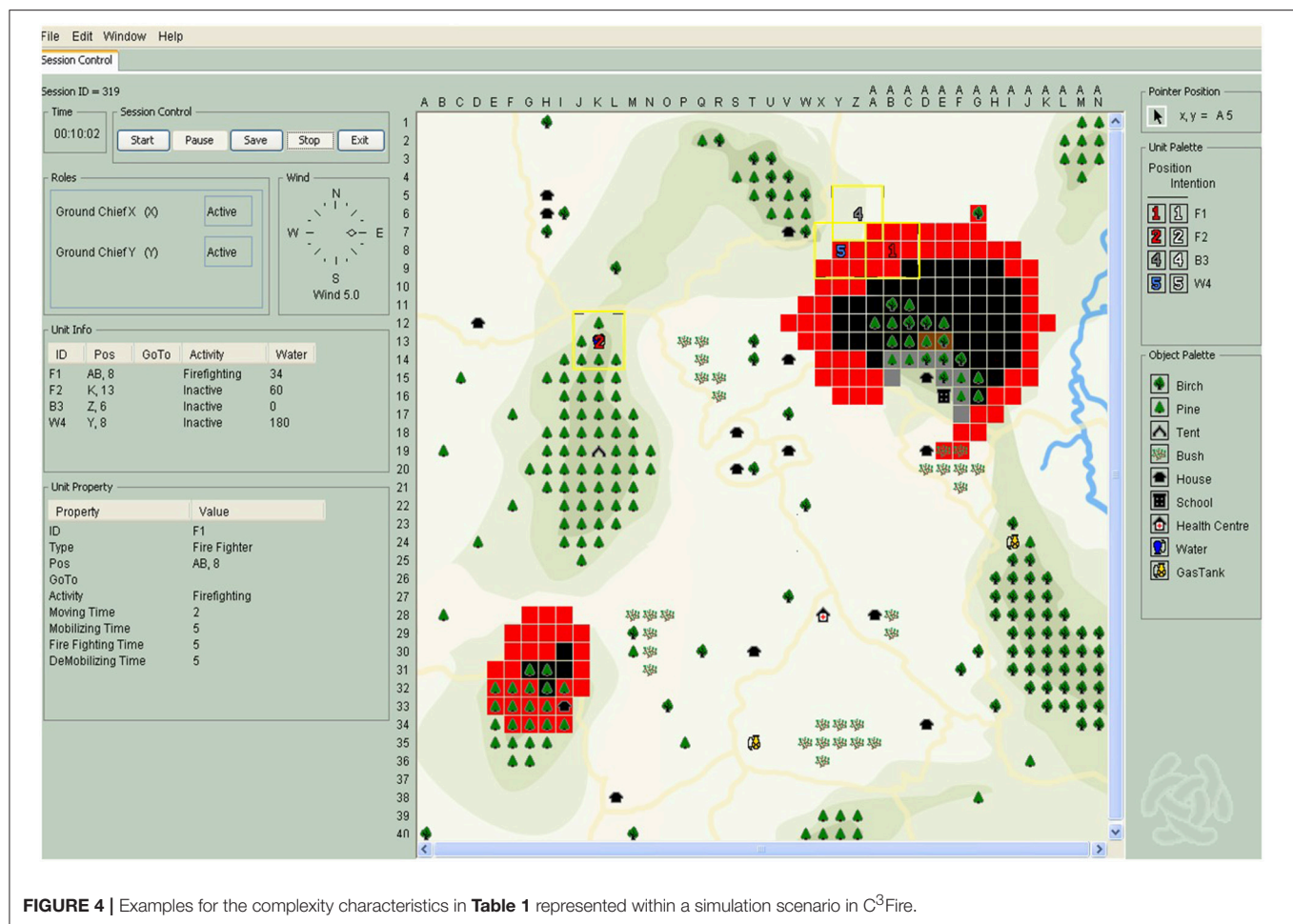


FIGURE 4 | Examples for the complexity characteristics in Table 1 represented within a simulation scenario in C³Fire.

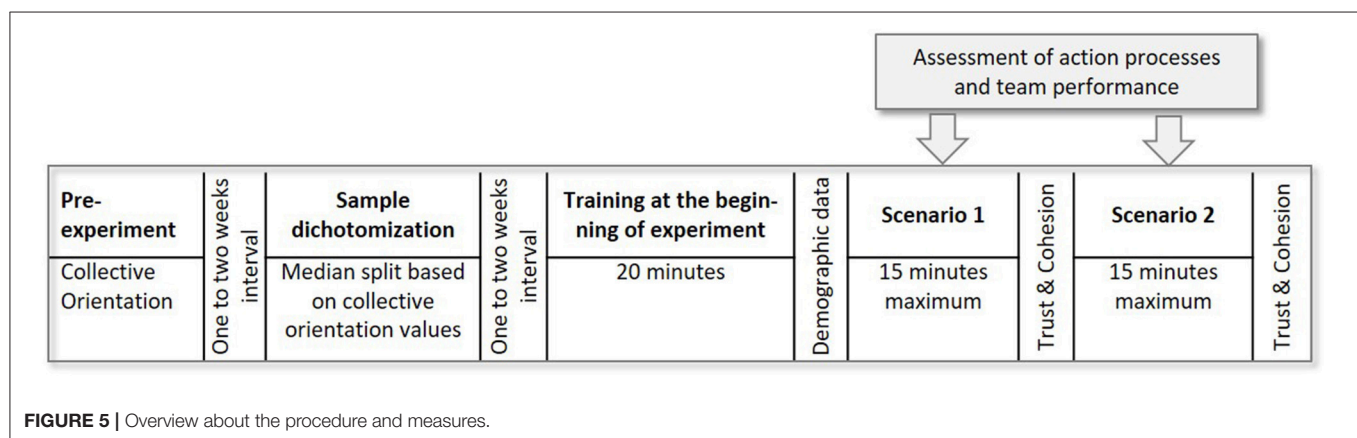


FIGURE 5 | Overview about the procedure and measures.

simulation scenario started and the participants had a maximum of 15 min to coordinate their actions to extinguish a forest fire whilst protecting houses and saving lives. After that, at measuring time T1, participants answered questionnaires assessing trust and cohesion within the team. Again, the teams worked on the following scenario 2 followed by a last round of questionnaires assessing trust and cohesion at T2.

## Measures

**Demographic data** such as age, sex, and study course were assessed after the training at the beginning of the experiment.

**Collective Orientation** was measured at an individual level with 16 items rated on a 5-point Likert scale (1 = *strongly disagree* to 5 = *strongly agree*) developed by the authors (Hagemann, 2017) based on the work of Driskell et al. (2010). The factorial

structure concerning the German-language CO scale was proven prior to this study ( $\chi^2 = 162.25$ ,  $df = 92$ ,  $p = 0.000$ ,  $\chi^2/df = 1.76$ , CFI = 0.97, TLI = 0.96, RMSEA = 0.040, CI = 0.030–0.051, SRMR = 0.043) and correlations for testing convergent and discriminant evidence of validity were satisfying. For example, CO correlated  $r = 0.09$  ( $p > 0.10$ ) with cohesion,  $r = 0.34$  ( $p < 0.01$ ) with cooperative interdependence and  $r = -0.28$  ( $p < 0.01$ ) with preference for solitude (Hagemann, 2017). An example item is “I find working on team projects to be very satisfying”. Coefficient alpha for this scale was 0.81.

**Trust** in team members’ integrity, trust in members’ task abilities and trust in members’ work-related attitudes (Geister et al., 2006) was measured with seven items rated on a 5-point Likert scale (1 = *strongly disagree* to 5 = *strongly agree*). An example item is “I can trust that I will have no additional demands due to lack of motivation of my team member.” Coefficient alpha for this scale was 0.83 (T1) and 0.87 (T2).

**Cohesion** was measured with a six-item scale from Riordan and Weatherly (1999) rated on a 5-point Likert scale (1 = *strongly disagree* to 5 = *strongly agree*). An example item is “In this team, there is a lot of team spirit among the members.” Coefficient alpha for this scale was 0.87 (T1) and 0.87 (T2).

### Action Process: Coordination

Successful coordination requires mechanisms that serve to manage dependencies between the teams’ activities and their resources. Coordination effectiveness was assessed based on the time the firefighting units spent without water in the field in relation to the total scenario time. This measure is an indicator of the effectiveness of resource-oriented coordination, as it reflects an efficient performance regarding the water refill process in C<sup>3</sup>Fire, which requires coordinated actions between the two firefighting units and one water tank unit (Lafond et al., 2011). The underlying assumption is that a more successful coordination process leads to fewer delays in conducting the refill process. Coordination was calculated by a formula and values ranged between 0 and 1, with lower values indicating better coordination in the team (see Jobidon et al., 2012).

$$\text{Coordination} = \frac{\text{time spent without water}}{\text{total time spent in scenario}}$$

### Team Performance

This measure related to the teams’ goals (limiting the number of burned out cells and saving as many houses/buildings as possible) and was quantified as the number of protected houses and the number of protected fields and bushes/trees in relation to the number of houses, fields, and bushes/trees, respectively, which would burn in a worst case scenario. This formula takes into account that teams needing more time for firefighting also have more burning cells and show a less successful performance than teams that are quick in firefighting. To determine the worst case scenario, both 15-min scenarios were run with no firefighting action taken. Thus, the particularities (e.g., how many houses would burn down if no action was taken) of each scenario were considered. Furthermore, the houses, bushes/trees and fields were weighted according to their differing importance, mirroring

the teams’ goals. Houses should be protected and were most important. Bushes/trees (middle importance) burn faster than fields (lowest importance) and foster the expansion of the fire. Values regarding team performance ranged between 0 and 7.99, with higher values indicating a better overall performance. Team performance was calculated as follows (see **Table 2**):

$$\text{TeamPerformance} = ((a/\max a)*5) + ((b/\max b)*2) + ((c/\max c)*1)$$

## RESULTS

Means, standard deviations, internal consistencies, and correlations for all study variables are provided in **Table 3**.

Team complex problem solving in scenario 1 correlated significantly negative with time without water in scenario 1, indicating that a high team performance is attended by the coordination behavior (as a team process). The same was true for scenario 2. In addition, time without water as an indicator for team coordination correlated significantly negative with the team members’ CO, indicating that team members with high CO values experience less time without water in the microworld than teams with members with low CO values.

In order to analyze the influence of CO on team process demands such as coordination processes and thereby performance within complex problem solving teams we tested whether CO would show an indirect effect on team performance through the teams’ coordination processes. To analyze this assumption, indirect effects in simple mediation models were estimated for both scenarios (see Preacher and Hayes, 2004). The mean for CO was 3.44 ( $SD = 0.32$ ) for teams with high CO values and it was 2.79 ( $SD = 0.35$ ) for teams with low CO values. The mean concerning team performance in scenario 1 for teams with high CO values was 6.30 ( $SD = 1.64$ ) and with low CO values 5.35 ( $SD = 2.30$ ). The mean concerning time without water (coordination behavior) for teams with high CO values was 0.16 ( $SD = 0.08$ ) and with low CO values 0.20 ( $SD = 0.09$ ). In scenario 2 the mean for team performance was

**TABLE 2 |** Explanation of formula for calculating team performance in both scenarios.

Symbol	Explanation
a	= number of protected houses (those that were not touched by fire)
b	= number of protected bushes/trees
c	= number of protected fields
max a	= number of affected houses in the worst case (those that are burned out, extinguished or still on fire)
max b	= number of affected bushes/trees in the worst case
max c	= number of affected fields in the worst case
5	= weighting of houses (highest priority)
2	= weighting of bushes/trees (middle priority)
1	= weighting of fields (lowest priority)



**TABLE 3 |** Means, standard deviations, internal consistencies, and correlations for all study variables.

	<i>M</i>	<i>SD</i>	$\alpha$	1	2	3	4	5	6	7	8
1 Performance scenario 1	5.82	2.03		1							
2 Performance scenario 2	5.31	2.53		0.31**	1						
3 Time without water scenario 1	0.177	0.09		−0.48**	−0.24**	1					
4 Time without water scenario 2	0.214	0.10		−0.02	−0.30**	0.25**	1				
5 Collective Orientation	3.12	0.46	0.81	0.14	0.20*	−0.20*	−0.42**	1			
6 Trust T1	4.43	0.51	0.83	0.18	0.06	−0.11	−0.08	0.05	1		
7 Trust T2	4.47	0.50	0.87	−0.02	0.06	−0.00	−0.12	−0.03	0.83**	1	
8 Cohesion T1	4.02	0.64	0.87	0.00	−0.09	−0.22*	−0.06	−0.17	0.47**	0.51**	1
9 Cohesion T2	4.01	0.65	0.87	0.01	−0.07	−0.17	−0.08	−0.18	0.39**	0.47**	0.87**

Performance range from 0 to 7.99; Time without Water range from 0 to 1 (lower values indicate a more effective handling of water); CO range from 1 to 5. \* $p < 0.05$ , \*\* $p < 0.01$ .

6.26 ( $SD = 2.51$ ) for teams with high CO values and it was 4.36 ( $SD = 2.24$ ) for teams with low CO values. The mean concerning time without water for teams with high CO values was 0.18 ( $SD = 0.08$ ) and with low CO values 0.25 ( $SD = 0.11$ ).

For analyzing indirect effects, CO was the independent variable, time without water the mediator and team performance the dependent variable. The findings indicated that CO has an indirect effect on team performance mediated by time without water for scenario 1 (Table 4) and scenario 2 (Table 5). In scenario 1, CO had no direct effect on team performance ( $b(YX)$ ), but CO significantly predicted time without water ( $b(MX)$ ). A significant total effect ( $b(YX)$ ) is not an assumption in the assessment of indirect effects, and therefore the non-significance of this relationship does not violate the analysis (see Preacher and Hayes, 2004, p. 719). Furthermore, time without water significantly predicted team performance when controlling for CO ( $b(YM.X)$ ), whereas the effect of CO on team performance was not significant when controlling for time without water ( $b(YX.M)$ ). The indirect effect was 0.40 and significant when using normal distribution and estimated with the Sobel test ( $z = 1.97$ ,  $p < 0.05$ ). The bootstrap procedure was applied to estimate the effect size not based on the assumption of normal distribution. As displayed in Table 4, the bootstrapped estimate of the indirect effect was 0.41 and the true indirect effect was estimated to lie between 0.0084 and 0.9215 with a 95% confidence interval. As zero is not in the 95% confidence interval, it can be concluded that the indirect effect is indeed significantly different from zero at  $p < 0.05$  (two-tailed).

Regarding scenario 2, CO had a direct effect on team performance ( $b(YX)$ ) and on time without water ( $b(MX)$ ). Again, time without water significantly predicted team performance when controlling for CO ( $b(YM.X)$ ), whereas the effect of CO on team performance was not significant when controlling for time without water ( $b(YX.M)$ ). This time, the indirect effect was 0.60 (Sobel test,  $z = 2.31$ ,  $p < 0.05$ ). As displayed in Table 5, the bootstrapped estimate of the indirect effect was 0.61 and the true indirect effect was estimated to lie between 0.1876 and 1.1014 with a 95% confidence interval and between 0.0340 and 1.2578 with a 99% confidence interval. Because zero is not in the 99% confidence interval, it can be concluded that the indirect effect is indeed significantly different from zero at  $p < 0.01$  (two-tailed).

**TABLE 4 |** Indirect Effect for Coordination and Team Performance in Scenario 1.

Effects	Coefficient	SE	T-ratio
b (YX)	00.5921	0.4047	1.4630
b (MX)	−00.0365	0.0171	−2.1329*
b (YM.X)	−10.9712	1.9735	−5.5592**
b (YX.M)	00.1920	0.3673	0.5228

#### Indirect Effect and Significance Using Normal Distribution

	Value	SE	LL 95 CI	UL 95 CI	Z
Sobel	0.4000	0.2037	0.0008	0.7993	1.9693*

#### Bootstrap Results for Indirect Effect

	Mean	SE	LL 95 CI	UL 95 CI	LL 99 CI	UL 99 CI
Effect	0.4134	0.2346	0.0084	0.9215	−0.0924	1.0999

Y = Team Performance Scenario 1; X = Collective Orientation T0; M = Coordination (time without water in scenario 1); Number of Bootstrap Resamples 5000. \* $p < 0.05$ , \*\* $p < 0.01$ .

The indirect effects for both scenarios are visualized in Figure 6. Summing up, the results support hypothesis 1 and indicate that CO has an indirect effect on team performance mediated by the teams' coordination behavior, an action process. That means, fulfilling team process demands affect the dynamic decision making quality of teams acting in complex situations and input variables such as CO influence the action processes within teams positively.

Trust between team members assessed after scenario 1 (T1) and after scenario 2 (T2) did not show any significant correlation with the coordination behavior or with team complex problem solving in scenarios 1 and 2 (Table 3). Thus, hypotheses 2a and 2b are not supported. Cohesion at T1 showed no significant relationship with team performance in both scenarios, one significant negative correlation ( $r = -0.22$ ,  $p < 0.05$ ) with the coordination behavior in scenario 1 and no correlation with the coordination behavior in scenario 2. Cohesion at T2 did not show any significant correlation with the coordination behavior or with

**TABLE 5 |** Indirect Effect for Coordination and Team Performance in Scenario 2.

Effects	Coefficient		SE		T-ratio
b (YX)	1.1086		0.4999		2.2176*
b (MX)	−0.0915		0.0185		−4.9419**
b (YM.X)	−6.5735		2.4634		−2.6685**
b (YX.M)	0.5071		0.5366		0.9450

Indirect Effect and Significance Using Normal Distribution					
	Value	SE	LL 95 CI	UL 95 CI	Z
Sobel	0.6015	0.2602	0.0915	1.1115	2.3117*

Bootstrap Results for Indirect Effect						
	Mean	SE	LL 95 CI	UL 95 CI	LL 99 CI	UL 99 CI
Effect	0.6055	0.2324	0.1876	1.1014	0.0340	1.2578

Y = Team Performance Scenario 2; X = Collective Orientation T0; M = Coordination (time without water in scenario 2); Number of Bootstrap Resamples 5000. \* $p < 0.05$ , \*\* $p < 0.01$ .

team performance in both scenarios. Thus, hypotheses 3a and 3b could also not be supported. Furthermore, the results showed no significant relations between CO and trust and cohesion. The correlations between trust and cohesion ranged between  $r = 0.39$  and  $r = 0.51$  ( $p < 0.01$ ).

## DISCUSSION

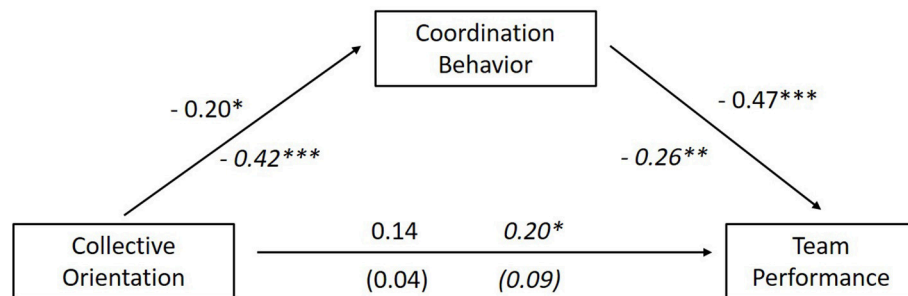
The purpose of our paper was first to give a sound theoretical overview and to combine theoretical approaches about team competencies and team process demands in collaborative complex problem solving and second to demonstrate the importance of selected team competencies and processes on team performance in complex problem solving by means of results from a laboratory study. We introduced the model of an idealized teamwork process that complex problem solving team pass through and integrated the relevant teamwork skills for interdependently working teams into it. Moreover, we highlighted the episodic aspect concerning complex problem solving in teams and combined the well-known transition, action, interpersonal and learning processes of teamwork with the idealized teamwork process model. Finally, we investigated the influence of trust, cohesion, and CO on action processes, such as coordination behavior of complex problem solving teams and on team performance.

Regarding hypothesis 1, studies have indicated that teams whose members have high CO values are more successful in their coordination processes and task accomplishment (Eby and Dobbins, 1997; Driskell et al., 2010; Hagemann, 2017), which may enhance team performance through considering task inputs from other team members, information sharing and strategizing (Salas et al., 2005). Thus, we had a close look on CO as an emergent state in the present study, because emergent states support the execution of behavioral processes. In order to analyze this indirect effect of CO on team performance via coordination

processes, we used the time, which firefighters spent without water in a scenario, as an indicator for high-quality coordination within the team. A small amount of time without water represents sharing information and resources between team members in a reciprocal manner, which are essential qualities of effective coordination (Ellington and Dierdorff, 2014). One of the two team members was in charge of the mobile water tank unit and therefore responsible for filling up the water tanks of his/her own firefighting unit and that of the other team member on time. In order to avoid running out of water for firefighting, the team members had to exchange information about, for example, their firefighting units' current and future positions in the field, their water levels, their strategies for extinguishing one or two fires, and the water tank unit's current and future position in the field. The simple mediation models showed that CO has an indirect effect on team performance mediated by time without water, supporting hypothesis 1. Thus, CO facilitates high-quality coordination within complex problem solving teams and this in turn influences decision-making and team performance positively (cf. Figure 1). These results support previous findings concerning the relationships between emergent states, such as CO, and the team process, such as action processes like coordination (Cannon-Bowers et al., 1995; Driskell et al., 2010) and between the team process and the team performance (Stevens and Campion, 1994; Dierdorff et al., 2011).

Hypotheses 2 and 3 analyzed the relationships between trust and cohesion and coordination and team performance. Because no correlations between trust and cohesion and the coordination behavior and team complex problem solving existed, further analyses, like mediation analyses, were unnecessary. In contrast to other studies (McAllister, 1995; Beal et al., 2003; Salas et al., 2005; Kozlowski and Ilgen, 2006), the present study was not able to detect effects of trust and cohesion on team processes, like action processes, or on team performance. This can be attributed to the restricted sample composition or the rather small sample size. Nevertheless, effect sizes were small to medium, so that they would have become significant with an increased sample sizes. The prerequisite, mentioned by the authors, that interdependence of the teamwork is important for identifying those effects, was given in the present study. Therefore, this aspect could not have been the reason for finding no effects concerning trust and cohesion. Trust and cohesion within the teams developed during working on the simulation scenarios while fighting fires, showed significant correlations with each other, and were unrelated to CO, which showed an effect on the coordination behavior and the team performance indeed. The results seem to implicate, that the influence of CO on action processes and team performance might be much more stronger than those of trust and cohesion. If these results can be replicated should be analyzed in future studies.

As the interdependent complex problem-solving task was a computer-based simulation, the results might have been affected by the participants' attitudes to using a computer. For example, computer affinity seems to be able to minimize potential fear of working with a simulation environment and might therefore, be able to contribute to successful performance in a computer-based team task. Although computers and other electronic



**FIGURE 6 |** Indirect effect of collective orientation on team performance via coordination within the teams for scenario 1 and 2, \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ , numbers in italic represent results from scenario 2, non-italic numbers are from scenario 1.

devices are pervasive in present-day life, computer aversion has to be considered in future studies within complex problem-solving research when applying computer-based simulation team tasks. As all of the participants were studying applied cognitive science, which is a mix of psychology and computer science, this problem might not have been influenced the present results. However, the specific composition of the sample reduces the external validity of the study and the generalizability of the results. A further limitation is the small sample size, so that moderate to small effects are difficult to detect.

Furthermore, laboratory research of teamwork might have certain limitations. Teamwork as demonstrated in this study fails to account for the fact that teams are not simple, static and isolated entities (McGrath et al., 2000). The validity of the results could be reduced insofar as the complex relationships in teams were not represented, the teamwork context was not considered, not all teammates and teams were comparable, and the characteristic as a dynamic system with a team history and future was not given in the present study. This could be a possible explanation why no effects of trust and cohesion were found in the present study. Maybe, the teams need more time working together on the simulation scenarios in order to show that trust and cohesion influence the coordination with the team and the team performance. Furthermore, Bell (2007) demonstrated in her meta-analysis that the relationship between team members' attitudes and the team's performance was proven more strongly in the field compared to the laboratory. In consideration of this fact, the findings of the present study concerning CO are remarkable and the simulation based microworld C<sup>3</sup>Fire (Granlund et al., 2001; Granlund, 2003) seems to be appropriate for analyzing complex problem solving in interdependently working teams.

An asset of the present study is, that the teams' action processes, the coordination performance, was assessed objectively based on logged data and was not a subjective measure, as is often the case in group and team research studies (cf. Van de Ven et al., 1976; Antoni and Hertel, 2009; Dierdorff et al., 2011; Ellington and Dierdorff, 2014). As coordination was the mediator in the analysis, this objective measurement supports the validity of the results.

## Outlook

As no transition processes such as mission analysis, formulation, and planning (Prince and Salas, 1993), goal specification (Prussia and Kinicki, 1996), and strategy formulation (Prince and Salas, 1993; Cannon-Bowers et al., 1995) as well as action processes such as monitoring progress toward goals (Cannon-Bowers et al., 1995) and systems monitoring (Fleishman and Zaccaro, 1992) were analyzed within the present study, future studies should collect data concerning these processes in order to show their importance on performance within complex problem solving teams. Because these processes are difficult to observe, subjective measurements are needed, for example asking the participants after each scenario how they have prioritized various tasks, if and when they have changed their strategy concerning protecting houses or fighting fires, and on which data within the scenarios they focused for collecting information for goal and systems monitoring. Another possibility could be using eye-tracking methods in order to collect data about collecting information for monitoring progress toward goals, e.g., collecting information how many cells are still burning, and systems monitoring, e.g., tracking team resources like water for firefighting.

CO is an emergent state and emergent states can be influenced by experience or learning, for example (Kozlowski and Ilgen, 2006). Learning processes (Edmondson, 1999), that Schmutz et al. (2016) added to the taxonomy of team processes developed by Marks et al. (2001) and which occur during transition and action phases and contribute to team effectiveness include e.g., *feedback*. Feedback can be useful for team learning when team learning is seen as a form of information processing (Hinsz et al., 1997). Because CO supports action processes, such as coordination and it can be influenced by learning, learning opportunities, such as feedback, seem to be important for successful task accomplishment and for supporting teams in handling complex situations or problems. If the team is temporarily and interpersonally unstable, as it is the case for most of the disaster or crisis management teams dealing with complex problems, there might be less opportunities for generating shared mental models by experiencing repetitive cycles of joint action (cf. Figure 2) and strategies such as cross training (Salas et al., 2007) or feedback might become more and more important for successful complex problem solving in teams. Thus, for future

research it would be of interest to analyze what kind of feedback is able to influence CO positively and therefore is able to enhance coordination and performance within complex problem-solving teams.

Depending on the type of feedback, different main points will be focused during the feedback (see Gabelica et al., 2012). Feedback can be differentiated into performance and process feedback. Process feedback can be further divided into task-related and interpersonal feedback. Besides these aspects, feedback can be given on a team-level or an individual-level. Combinations of the various kinds of feedback are possible and are analyzed in research concerning their influence on e.g., self- and team-regulatory processes and team performance (Prussia and Kinicki, 1996; Hinsz et al., 1997; Jung and Sosik, 2003; Gabelica et al., 2012). For future studies it would be relevant to analyze, whether it is possible to positively influence the CO of team members and therefore action processes such as coordination and team performance or not. A focus could be on the learning processes, especially on feedback, and its influence on CO in complex problem solving teams. So far, no studies exist that analyzed the relationship between feedback and a change in CO, even though researchers already discuss the possibility that team-level process feedback shifts attention processes on team actions and team learning (McLeod et al., 1992; Hinsz et al., 1997). These results would be very helpful for training programs for fire service or police or medical teams working in complex environments and solving problems collaboratively, in order to support their team working and their performance.

In summary, the idealized teamwork process model is in combination with the transition, action, interpersonal and learning processes a good framework for analyzing the impact of teamwork competencies and teamwork processes in detail on team performance in complex environments. Overall, the framework offers further possibilities for investigating the

influence of teamwork competencies on diverse processes and teamwork outcomes in complex problem solving teams than demonstrated here. The results of our study provide evidence of how CO influences complex problem solving teams and their performance. Accordingly, future researchers and practitioners would be well advised to find interventions how to influence CO and support interdependently working teams.

## ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Ethical guidelines of the German Association of Psychology, Ethics committee of the University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Ethics committee of the University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science.

## AUTHOR CONTRIBUTIONS

VH and AK were responsible for the conception of the work and the study design. VH analyzed and interpreted the collected data. VH and AK drafted the manuscript. They approved it for publication and act as guarantors for the overall content.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fpsyg.2017.01730/full#supplementary-material>

## REFERENCES

- Antoni, C., and Hertel, G. (2009). Team processes, their antecedents and consequences: implications for different types of teamwork. *Eur. J. Work Organ. Psychol.* 18, 253–266. doi: 10.1080/13594320802095502
- Arthur, W., Edwards, B., Bell, S., Villado, A., and Bennis, W. (2005). Team task analysis: identifying tasks and jobs that are team based. *Hum. Factors* 47, 654–669. doi: 10.1518/001872005774860087
- Badke-Schaub, P. (2008). Teamarbeit und Teamführung: Erfolgsfaktoren und sicheres Handeln. [Teamwork and Team leadership: Factors of success and reliable action], in *Führung und Teamarbeit in kritischen Situationen [Leadership and teamwork in critical situations]* eds C. Buerschaper and S. Starke (Frankfurt: Verlag für Polizeiwissenschaft), 3–19.
- Beal, D. J., Cohen, R. R., Burke, M. J., and McLendon, C. L. (2003). Cohesion and performance in groups: a meta-analytic clarification of construct relations. *J. Appl. Psychol.* 88, 989–1004. doi: 10.1037/0021-9010.88.6.989
- Bell, S. T. (2007). Deep-level composition variables as predictors of team performance: a meta-analysis. *J. Appl. Psychol.* 92, 595–615. doi: 10.1037/0021-9010.92.3.595
- Bower, G. H., and Hilgard, E. R. (1981). *Theories of Learning*. Englewood Cliffs, NJ: Prentice-Hall.
- Bredenkamp, J. (1998). *Lernen, Erinnern, Vergessen [Learning, Remembering, Forgetting]*. München: C.H. Beck.
- Brehmer, B. (1992). Dynamic decision-making: human control of complex systems. *Acta Psychol.* 81, 211–241. doi: 10.1016/0001-6918(92)90019-A
- Brehmer, B., and Dörner, D. (1993). Experiments with computer-simulated microworlds: escaping both the narrow straits of the laboratory and the deep blue sea of the field study. *Comput. Hum. Behav.* 9, 171–184. doi: 10.1016/0747-5632(93)90005-D
- Campion, M. A., Medsker, G. J., and Higgs, C. (1993). Relations between work group characteristics and effectiveness: implications for designing effective work groups. *Pers. Psychol.* 46, 823–850. doi: 10.1111/j.1744-6570.1993.tb01571.x
- Cannon-Bowers, J. A., Salas, E., and Converse, S. (1993). “Shared mental models in expert team decision making,” in *Individual and Group Decision Making*, ed N. J. Castellan (Hillsdale, NJ: Lawrence Erlbaum Associates), 221–246.
- Cannon-Bowers, J. A., Tannenbaum, S. I., Salas, E., and Volpe, C. E. (1995). “Defining competencies and establishing team training requirements,” in *Team Effectiveness and Decision Making in Organizations*, eds R. A. Guzzo and E. Salas and Associates (San Francisco, CA: Jossey-Bass), 333–380.
- DeChurch, L. A., and Mesmer-Magnus, J. R. (2010). The cognitive underpinnings of effective teamwork: a meta-analysis. *J. Appl. Psychol.* 95, 32–53. doi: 10.1037/a0017328
- Dierdorff, E. C., Bell, S. T., and Belohlav, J. A. (2011). The “power of we”: effects of psychological collectivism on team performance over time. *J. Appl. Psychol.* 96, 247–262. doi: 10.1037/a0020929



- Dörner, D. (1989/2003). *Die Logik des Misslingens. Strategisches Denken in komplexen Situationen [The logic of failure. Strategic thinking in complex situations]* 11th Edn. Reinbeck: rororo.
- Dörner, D., Kreuzig, H. W., Reither, F., and Stäudel, T. (1983). *Lohhausen. Vom Umgang mit Unbestimmtheit und Komplexität*. Bern; Stuttgart; Wien: Verlag Hans Huber.
- Driskell, J. E., and Salas, E. (1992). Collective behavior and team performance. *Hum. Factors* 34, 277–288. doi: 10.1177/001872089203400303
- Driskell, J., Salas, E., and Hughes, S. (2010). Collective orientation and team performance: development of an individual differences measure. *Hum. Factors* 52, 316–328. doi: 10.1177/0018720809359522
- Eby, L. T., and Dobbins, G. H. (1997). Collectivistic orientation in teams: an individual and group-level analysis. *J. Organ. Behav.* 18, 275–295. doi: 10.1002/(SICI)1099-1379(199705)18:3<275::AID-JOB796>3.0.CO;2-C
- Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Adm. Sci. Q.* 44, 350–383. doi: 10.2307/2666999
- Ellington, J. K., and Dierdorff, E. C. (2014). Individual learning in team training: self-regulation and team context effects. *Small Group Res.* 45, 37–67. doi: 10.1177/1046496413511670
- Endsley, M. R. (1999). “Situation Awareness in Aviation Systems,” in *Handbook of Aviation Human Factors*, eds D. J. Garland, J. A. Wise, and V. D. Hopkin (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 257–276.
- Endsley, M. R., and Robertson, M. M. (2000). “Training for Situation Awareness in Individuals and Teams,” in *Situation awareness Analysis and Measurement*, eds M. R. Endsley and D. J. Garland (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 349–365.
- Fiore, S. M., Rosen, M. A., Smith-Jentsch, K. A., Salas, E., Letsky, M., and Warner, N. (2010). Toward an understanding of macrocognition in teams: predicting processes in complex collaborative contexts. *Hum. Factors* 52, 203–224. doi: 10.1177/0018720810369807
- Fleishman, E. A., and Zaccaro, S. J. (1992). “Toward a taxonomy of team performance functions,” in *Teams: Their Training and Performance*, eds R. W. Swezey and E. Salas (Norwood, NJ: Ablex), 31–56.
- Flin, R., O’Connor, P., and Crichton, M. (2008). *Safety at the Sharp End*. Aldershot: Ashgate.
- Funke, J. (1995). “Experimental research on complex problem solving,” in *Complex Problem Solving: The European Perspective* eds P. A. Frensch and J. Funke (Hillsdale, NJ: Lawrence Erlbaum Associates), 243–268.
- Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Think. Reason.* 7, 69–89. doi: 10.1080/13546780042000046
- Funke, J. (2012). “Complex Problem Solving,” in *Encyclopedia of the Sciences of Learning* ed N. M. Seel (Heidelberg: Springer), 682–685.
- Gabelica, C., van den Bossche, P., de Maeyer, S., Segers, M., and Gijssels, W. (2014). The effect of team feedback and guided reflexivity on team performance change. *Learn. Instruct.* 34, 86–96. doi: 10.1016/j.learninstruct.2014.09.001
- Gabelica, C., Van den Bossche, P., Segers, M., and Gijssels, W. (2012). Feedback, a powerful lever in teams: a review. *Educ. Res. Rev.* 7, 123–144. doi: 10.1016/j.edurev.2011.11.003
- Geister, S., Konradt, U., and Hertel, G. (2006). Effects of process feedback on motivation, satisfaction, and performance in virtual teams. *Small Group Res.* 37, 459–489. doi: 10.1177/1046496406292337
- Gibson, C. B. (1999). Do they do what they believe they can? group efficacy and group effectiveness across tasks and cultures. *Acad. Manag. J.* 42, 138–152. doi: 10.2307/257089
- Granlund, R. (2003). Monitoring experiences from command and control research with the C<sup>3</sup>Fire microworld. *Cogn. Technol. Work* 5, 183–190. doi: 10.1007/s10111-003-0129-8
- Granlund, R., and Johansson, B. (2004). “Monitoring distributed collaboration in the C<sup>3</sup>Fire Microworld,” in *Scaled Worlds: Development, Validation and Applications*, eds G. Schiflett, L. R. Elliot, E. Salas, and M. D. Coovet (Aldershot: Ashgate), 37–48.
- Granlund, R., Johansson, B., and Persson, M. (2001). “C<sup>3</sup>Fire a micro-world for collaboration training and investigations in the ROLF environment,” in *Proceedings of 42nd Conference on Simulation and Modeling: Simulation in Theory and Practice* (Porsgrunn).
- Hackman, J. R. (1987). “The design of work teams,” in *Handbook of Organizational Behavior* ed J. W. Lorsch (Englewood Cliffs, NJ: Prentice-Hall), 315–342.
- Hagemann, V. (2011). *Trainingsentwicklung für High Responsibility Teams [Training development for High Responsibility Teams]*. Lengerich: Pabst Verlag.
- Hagemann, V. (2017). Development of a German-language questionnaire to measure collective orientation as an individual attitude. *Swiss J. Psychol.* 76, 91–105. doi: 10.1024/1421-0185/a000198
- Hagemann, V., Kluge, A., and Ritzmann, S. (2011). High responsibility teams - Eine systematische Analyse von Teamarbeitskontexten für einen effektiven Kompetenzerwerb [A systematic analysis of teamwork contexts for effective competence acquisition]. *Psychologie des Alltagshandelns* 4, 22–42. Available online at: [http://www.allgemeine-psychologie.info/cms/images/stories/allgpsy\\_journal/Vol%204%20No%201/hagemann\\_kluge\\_ritzmann.pdf](http://www.allgemeine-psychologie.info/cms/images/stories/allgpsy_journal/Vol%204%20No%201/hagemann_kluge_ritzmann.pdf)
- Hagemann, V., Kluge, A., and Ritzmann, S. (2012). Flexibility under complexity: work contexts, task profiles and team processes of high responsibility teams. *Empl. Relat.* 34, 322–338. doi: 10.1108/01425451211217734
- Hertel, G., Konradt, U., and Orlikowski, B. (2004). Managing distance by interdependence: goal setting, task interdependence, and team-based rewards in virtual teams. *Euro. J. Work Organ. Psychol.* 13, 1–28. doi: 10.1080/13594320344000228
- Hinsz, V., Tindale, R., and Vollrath, D. (1997). The emerging concept of groups as information processors. *Psychol. Bull.* 121, 43–64. doi: 10.1037/0033-2909.121.1.43
- Hollingshead, A. B., Gupta, N., Yoon, K., and Brandon, D. (2012). “Transactive memory theory and teams: past, present, and future,” in *Theories of Team Cognition*, eds E. Salas, S. M. Fiore and M. Letsky (New York, NY: Routledge Taylor & Francis Group), 421–455.
- Ilgen, D. R., Hollenbeck, J. R., Johnson, M., and Jundt, D. (2005). Teams in organizations: from input-process-output models to IMO models. *Annu. Rev. Psychol.* 56, 517–543. doi: 10.1146/annurev.psych.56.091103.070250
- Jobidon, M.-E., Muller-Gass, A., Duncan, M., and Blais, A.-R. (2012). The enhance of mental models and its impact on teamwork. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 56, 1703–1707. doi: 10.1177/1071181312561341
- Jung, D. I., and Sosik, J. J. (2003). Group potency and collective efficacy. *Group Organ. Manag.* 28, 366–391. doi: 10.1177/1059601102250821
- Kluge, A. (2014). *The Acquisition of Knowledge and Skills for Taskwork and Teamwork to Control Complex Technical Systems : A Cognitive and Macroergonomics Perspective*. Dordrecht: Springer.
- Kluge, A., Hagemann, V., and Ritzmann, S. (2014). “Military crew resource management – Das Streben nach der bestmöglichen Teamarbeit [Striving for the best of teamwork],” in *Psychologie für Einsatz und Notfall [Psychology for mission and emergency]*, eds G. Kreim, S. Bruns and B. Völker (Bonn: Bernard & Graefe in der Mönch Verlagsgesellschaft mbH), 141–152.
- Kluge, A., Sauer, J., Schüler, K., and Burkolder, D. (2009). Designing training for process control simulators: a review of empirical findings and current practices, theoretical issues in ergonomics *Science* 10, 489–509. doi: 10.1080/14639220902982192
- Kozlowski, S. W. J., and Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychol. Sci. Public Interest* 7, 77–124. doi: 10.1111/j.1529-1006.2006.00030.x
- Lafond, D., Jobidon, M.-E., Aubé, C., and Tremblay, S. (2011). Evidence of structure-specific teamwork requirements and implications for team design. *Small Group Res.* 42, 507–535. doi: 10.1177/1046496410397617
- Marks, M. A., Mathieu, J. E., and Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Acad. Manag. Rev.* 26, 356–376. doi: 10.2307/259182
- Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., and Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *J. Appl. Psychol.* 85, 273–283. doi: 10.1037/0021-9010.85.2.273
- McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Acad. Manag. J.* 38, 24–59. doi: 10.2307/256727
- McGrath, J. E. (1984). *Groups: Interaction and Performance*. Englewood Cliffs, NJ: Prentice-Hall.
- McGrath, J. E., Arrow, H., and Berdahl, J. L. (2000). The study of groups: past, present, and future. *Pers. Soc. Psychol. Rev.* 4, 95–105. doi: 10.1207/S15327957PSPR0401\_8
- McLeod, P. L., Liker, J. K., and Lobel, S. A. (1992). Process feedback in task groups: an application of goal setting. *J. Appl. Behav. Sci.* 28, 15–41. doi: 10.1177/0021886392281003

- Morgan, B. B., Salas, E., and Glickman, A. S. (1993). An analysis of team evolution and maturation. *J. Gen. Psychol.* 120, 277–291. doi: 10.1080/00221309.1993.9711148
- Ormerod, T. C., Richardson, J., and Shepherd, A. (1998). Enhancing the usability of a task analysis method: a notation and environment for requirements specification. *Ergonomics* 41, 1642–1663. doi: 10.1080/001401398186117
- Preacher, K., and Hayes, A. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behav. Res. Methods Instrum. Comput.* 36, 717–731. doi: 10.3758/BF03206553
- Prince, C., and Salas, E. (1993). “Training and research for teamwork in the military aircrew,” in *Cockpit Resource Management*, eds E. L. Wiener, B. G. Kanki, and R. L. Helmreich (San Diego, CA: Academic Press), 337–366.
- Prussia, G. E., and Kinicki, A. J. (1996). A motivation investigation of group effectiveness using social-cognitive theory. *J. Appl. Psychol.* 81, 187–198. doi: 10.1037/0021-9010.81.2.187
- Riordan, C. M., and Weatherly, E. W. (1999). Defining and measuring employees’ identification with their work groups. *Educ. Psychol. Meas.* 59, 310–324. doi: 10.1177/00131649921969866
- Roth, E. M., and Woods, D. D. (1988). Aiding human performance i: cognitive analysis. *Trav. Hum.* 51, 39–64.
- Salas, E., Cooke, N. J., and Rosen, M. A. (2008). On teams, teamwork, and team performance: discoveries and developments. *Hum. Factors* 50, 540–547. doi: 10.1518/001872008X288457
- Salas, E., Nichols, D. R., and Driskell, J. E. (2007). Testing three team training strategies in intact teams. *Small Group Res.* 38, 471–488. doi: 10.1177/1046496407304332
- Salas, E., Sims, D., and Burke, S. (2005). Is there a “big five” in teamwork? *Small Group Res.* 36, 555–599. doi: 10.1177/1046496405277134
- Schmutz, J., Welp, A., and Kolbe, M. (2016). “Teamwork in healthcare organizations,” in *Management Innovations for Health Care Organizations*, eds A. Örtengren, C. A. Löfström and R. Sheaff (New York, NY: Routledge Taylor & Francis), 359–377.
- Serfaty, D., Entin, E. E., and Johnston, J. H. (1998). “Team coordination training,” in *Making Decisions Under Stress*, eds J. A. Cannon-Bowers and E. Salas (Washington, DC: American Psychological Association), 221–246.
- Shea, G. P., and Guzzo, R. A. (1987). Group effectiveness: what really matters? *Sloan Manage. Rev.* 28, 25–31.
- Smith-Jentsch, K. A., Baker, D. P., Salas, E., and Cannon-Bowers, J. A. (2001). “Uncovering differences in team competency requirements: The case of air traffic control teams,” in *Improving Teamwork in Organizations. Applications of Resource Management Training*, eds E. Salas, C. A. Bowers, and E. Edens (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 31–54.
- Stajkovic, A. D., Lee, D., and Nyberg, A. J. (2009). Collective efficacy, group potency, and group performance: meta-analyses of their relationships, and test of a mediation model. *J. Appl. Psychol.* 94, 814–828. doi: 10.1037/a0015659
- Stevens, M. J., and Campion, M. A. (1994). The knowledge, skill, and ability requirements for teamwork: implications for human resource management. *J. Manage.* 20, 503–530. doi: 10.1177/014920639402000210
- Ulich, E. (1995). “Gestaltung von Arbeitstätigkeiten [Designing job tasks],” in *Lehrbuch Organisationspsychologie [Schoolbook Organizational Psychology]*, ed H. Schuler (Bern: Huber), 189–208.
- Van de Ven, A. H., Delbecq, A. L., and Koenig, R. (1976). Determinants of coordination modes with organizations. *Am. Sociol. Rev.* 41, 322–338. doi: 10.2307/209477
- Wageman, R. (1995). Interdependence and group effectiveness. *Adm. Sci. Q.* 40, 145–180. doi: 10.2307/2393703
- Waller, M. J., Gupta, N., and Giambattista, R. C. (2004). Effects of adaptive behaviors and shared mental models on control crew performance. *Manage. Sci.* 50, 1534–1544. doi: 10.1287/mnsc.1040.0210
- Wilson, K. A., Salas, E., and Andrews, D. H. (2010). “Preventing errors in the heat of the battle: formal and informal learning strategies to prevent teamwork breakdowns,” in *Human Factors Issues in Combat Identification*, eds D. H. Andrews, R. P. Herz, and M. B. Wolf (Aldershot: Ashgate), 1–28.
- Zsombok, C. E. (1997). “Naturalistic decision making: where are we now?,” in *Naturalistic Decision Making*, eds C. E. Zsombok and G. Klein (New York, NY: Routledge), 3–16.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Hagemann and Kluge. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Beyond Psychometrics: The Difference between Difficult Problem Solving and Complex Problem Solving

Jens F. Beckmann<sup>1\*</sup>, Damian P. Birney<sup>2</sup> and Natassia Goode<sup>3</sup>

<sup>1</sup> School of Education, Durham University, Durham, United Kingdom, <sup>2</sup> School of Psychology, University of Sydney, Sydney, NSW, Australia, <sup>3</sup> Centre for Human Factors and Sociotechnical Systems, University of the Sunshine Coast, Sunshine Coast, QLD, Australia

## OPEN ACCESS

### Edited by:

Wolfgang Schoppek,  
University of Bayreuth, Germany

### Reviewed by:

Andreas Fischer,  
Forschungsinstitut Betriebliche  
Bildung (f-bb), Germany  
Manuel Bedia,  
University of Zaragoza, Spain

### \*Correspondence:

Jens F. Beckmann  
j.beckmann@durham.ac.uk

### Specialty section:

This article was submitted to  
Cognitive Science,  
a section of the journal  
Frontiers in Psychology

**Received:** 26 June 2017

**Accepted:** 20 September 2017

**Published:** 10 October 2017

### Citation:

Beckmann JF, Birney DP and  
Goode N (2017) Beyond  
Psychometrics: The Difference  
between Difficult Problem Solving  
and Complex Problem Solving.  
Front. Psychol. 8:1739.  
doi: 10.3389/fpsyg.2017.01739

In this paper we argue that a synthesis of findings across the various sub-areas of research in complex problem solving and consequently progress in theory building is hampered by an insufficient differentiation of complexity and difficulty. In the proposed framework of person, task, and situation (PTS), complexity is conceptualized as a quality that is determined by the cognitive demands that the characteristics of the task and the situation impose. Difficulty represents the quantifiable level of a person's success in dealing with such demands. We use the well-documented "semantic effect" as an exemplar for testing some of the conceptual assumptions derived from the PTS framework. We demonstrate how a differentiation between complexity and difficulty can help take beyond a potentially too narrowly defined psychometric perspective and subsequently gain a better understanding of the cognitive mechanisms behind this effect. In an empirical study a total of 240 university students were randomly allocated to one of four conditions. The four conditions resulted from contrasting the semanticity level of the variable labels used in the CPS system (high vs. low) and two instruction conditions for how to explore the CPS system's causal structure (starting with the assumption that all relationships between variables existed vs. starting with the assumption that none of the relationships existed). The variation in the instruction aimed at inducing knowledge acquisition processes of either (1) systematic elimination of presumptions, or (2) systematic compilation of a mental representation of the causal structure underpinning the system. Results indicate that (a) it is more complex to adopt a "blank slate" perspective under high semanticity as it requires processes of inhibiting prior assumptions, and (b) it seems more difficult to employ a systematic heuristic when testing against presumptions. In combination, situational characteristics, such as the semanticity of variable labels, have the potential to trigger qualitatively different tasks. Failing to differentiate between 'task' and 'situation' as independent sources of complexity and treating complexity and difficulty synonymously threaten the validity of performance scores obtained in CPS research.

**Keywords:** complex problem solving, semantic effect, complexity vs. difficulty, systematicity, person-task-situation

## INTRODUCTION

Complex problem solving (CPS) is an umbrella term for a diverse range of approaches to research, learning and assessment. A common denominator of all these approaches is the use of a computerized simulation of some abstract or contextualized system. This is where the main commonality ends. However, when considering the various different ways problem solvers can interact with these simulations and the wide variety of different purposes of their use, it becomes apparent that the term CPS has many different meanings. One meaning refers to a *research paradigm* that aims to study “complex” cognition in the context of information processing, decision-making, causal reasoning, or learning (Beckmann, 1994; Beckmann and Guthke, 1995; Frensch and Funke, 1995; Guthke et al., 1995). In other domains (e.g., Greiff et al., 2015), CPS has been considered as an *ability-related construct* (or set of constructs). One example is the ability to deal with uncertainty (e.g., Osman, 2010) with its conceptual – yet not always empirically aligned – links to reasoning and (fluid) intelligence (Funke and Frensch, 2007; Stadler et al., 2015). CPS has also started to establish its use in relation to an assessment approach, be it in smaller-scale studies in relation to personnel decisions (Wood et al., 2009) or in relation to larger-scale educational attainment assessment exercises such as PISA (OECD, 2013; Funke et al., 2017). Within an assessment context, CPS is often discussed as a *skill or competency* (rather than ability). On the one hand, the shared use of the term CPS in these contexts tends to belie the conceptual, and concomitantly, methodological diversity in this field of research; on the other hand, such diversity in meaning raises the suspicion of an insufficient conceptual foundation of CPS.

As a look beyond CPS and at scientific theory building paradigms generally reminds us, a lack of conceptual grounding tends to result in definitions for the respective target constructs that are predominantly operational (rather than conceptual). This is evident in the CPS literature, as CPS has often been used as a descriptor of the kind of *behavior* observable when individuals are confronted with a specific kind of challenge (i.e., CPS is what problem solvers do when dealing with complex problems). As a corollary of a preponderance of operational definitions, research and subsequent publications seem to be heavily focussing on psychometric characteristics of CPS simulations as measurement tools. In its extreme, such a situation might be perceived as delegating conceptual decisions to statistical procedures.

With this paper, we aim to go beyond the psychometrically driven approach to CPS and to contribute to a more theory-based positioning of it within a nomological network of cognition. Our argumentation leads to an empirical investigation that explicitly differentiates manipulations of complexity (the conceptual) from experiences of difficulty (the psychometric) and in so doing, demonstrate the importance of separating statistical and conceptual issues in the investigation of CPS.

## COMPLEXITY vs. DIFFICULTY

One symptom of a predominantly psychometric view on CPS is the lack of a distinction between complexity and difficulty (Beckmann and Goode, 2017). Difficulty is a psychometric concept with rather limited *explanatory* value. In general terms, difficulty provides a *descriptive* account of some items being answered correctly by a smaller proportion of individuals than other items, thus creating the basis for them being labeled as more difficult. When interested in the reasons for their higher levels of difficulty, one is confronted with a tautological reference (common to classical-test-theory) to the lower proportion of correct responses these items tend to attract. Actual explanations as to why this might be the case, however, need to go beyond such circularity. An analysis of the cognitive behavior required to tackle the problem posed by an item, as well as a reflection of the circumstances under which the item is expected to be solved, feeds into the notion of complexity. In this regard, complexity reflects *ex ante* considerations of the cognitive demands imposed by the task and the circumstances under which the task is to be performed (i.e., predictions), which makes complexity a primarily *cognitive concept*. Difficulty is experiential, person-bound and by definition, statistical. It is a reflection of how well individuals (with their individual differences in ability, knowledge, skills, motivation, etc.) deal with complexity, which makes it a *psychometric concept*.

At a first and rather pragmatic glance, such distinction may seem pedantic. After all, so it could be claimed, the presumptions linking difficulty (statistics) with complexity (theory) are built on a wealth of replicable scientific evidence. Therefore, so one might argue, our criticism would be considered not only unfounded in practice, but even counterproductive to the pursuit of knowledge. However, it is very easy to demonstrate this is not the case, and that when considered through person–task–situation (PTS) interactions, the broader CPS paradigm proves to be particularly in need of such a distinction. In the following we present a framework that allows for a conceptual differentiation between complexity and difficulty in the context of CPS. We then empirically test core arguments derived from this framework.

We start by taking the perspective of CPS as a research paradigm that utilizes computerized scenarios as task stimuli. These computerized scenarios or microworlds can be conceptualized as systems (e.g., Funke, 1985, 1992). In their simplest form, such systems comprise two kinds of variables that are causally linked, input variables and output variables and the interconnectedness of these system variables can be algorithmically described through linear structural equations. Such systems are considered “dynamic” if output variables change both as an effect of problem solvers inputs and independently over time.

In the contexts of research, assessment and learning, problem solvers are usually asked to first explore the unknown causal structure of these systems. In general, such an *exploration phase* serves the purpose of knowledge acquisition. In a subsequent *control phase*, problem solvers are then asked to reach and maintain pre-defined goal states in the output variables. The objective here is the application or utilization of the knowledge



acquired during the exploration phase. In the terminology of generic problem solving, the typical CPS constellation is where a particular set of operations has to be identified (i.e., knowledge acquired) that will bring the system from a given initial state to a set target one (i.e., system control).

## CPS IN THE THREE DIMENSIONAL SPACE OF PERSON, TASK, AND SITUATION (PTS)

As has been discussed previously (Beckmann, 2010; Birney et al., 2016; Beckmann and Goode, 2017), psychological research takes place in the three dimensional space of Person, Task, and Situation variables. The definition of task has two sub-facets, the *task qua task* and the *task as behavior requirement* (McGrath and Altman, 1966; Hackman, 1969; Wood, 1986). The *task qua task* facet refers to the physical characteristics of the stimuli the problem solver is confronted with. In the context of CPS, these are the characteristics of the CPS scenario which include, but are not limited to, the number of variables or the density of their interconnectedness. The *task as behavior requirement* refers to what the problem solver is instructed to do. In the context of CPS, this could be, for example, to either freely interact with the given system to uncover its causal structure, or to control this system, i.e., to reach and maintain a set of target states in the output variables. Both, task qua task and task as behavior requirements contribute interactively to the complexity of the CPS task. That is, being confronted with the same system (*task qua task*) but with different instructions – as communicated *task as behavior requirement* – results in different tasks with different levels of complexity. In short, different tasks require different sets of *expected*<sup>1</sup> (cognitive) behavior and therefore contribute differently to complexity.

The definition of *situation* refers to the environment or the circumstances in which a given task is to be performed (“task environment” as described by Newell and Simon, 1972, p. 55). In the context of CPS this includes situational characteristics such as whether a causal diagram (i.e., a graphical representation of the causal structure) is available or not when being asked to control the system. Knowing or being able to anticipate the target states during the exploration phase (or not) would be another situational characteristic. As these and other variations in circumstantial characteristics are also expected to result in differing sets of cognitive behaviors (despite being confronted with the same system and the same instruction), the situation is conceptualized as another contributor to complexity.

So far we have identified that both the task (with its two facets *task qua task* and *task as behavior requirement*) and the situation contribute to complexity. The third category of variables is linked to *Person* and includes, inter alia, individual differences in reasoning ability, information processing capacity, motivation, working memory, experience, and knowledge. Observed

performance is the resultant of the *difficulty* individuals’ experience in dealing with the complexity imposed by the task and the situation. In short, difficulty is the observable, subjective reflection of complexity.

Experimental research in psychology, irrespective of its focus, builds on observing variation in one component of this tripartite system of variables (i.e., Person, Task, and Situation) whilst the variation in the other two is either controlled for or, more or less systematically manipulated. The dominant experimental paradigms can be defined by their focus on one of these three components. For instance, in an *assessment* context, test takers are confronted with a standardized set of tasks under standardized instructions (e.g., to control a particular microworld) and in standardized situations (e.g., after a knowledge acquisition phase that resulted in a causal diagram, which is made available on the computer screen). Standardization ensures that all test takers are dealing with the same level of complexity (as it is defined by the system, the instructed task and the situation), so that observed variability in performance scores between (and occasionally within) individuals can be attributed to individual differences in conceptually relevant person characteristics (e.g., reasoning ability).

In comparison, in the context of *cognition research*, participants are confronted with systems (task qua task) or situations that differ systematically as part of experimental manipulations. Randomization in the allocation of participants to conditions aims at controlling for potential effects of individual differences. This allows for observed variability in average performances scores across conditions to be attributed to differences in complexity caused by the variation in task and/or situation characteristics.

In the context of *instructional design* research, as another example, the situational features of a (learning) task are systematically varied (e.g., availability and location of information – say, 0, 1, or more mouse-clicks away) whilst the task as behavior requirements (e.g., to acquire structural knowledge) and the task qua task (i.e., the system) are kept the same across learners. Observed performance differences are then interpreted as indications of how various *situational* variables (e.g., interface features) make a learning task more or less complex.

In sum, complexity and difficulty are different concepts. Failing to differentiate between the two is problematic in at least two ways. First, equating (observation-based and psychometrically derived) difficulty with complexity serves to perpetuate the circular argument of that what is difficult must be complex, and what makes something complex is its difficulty. Second, equating (task and situation analysis based) complexity with difficulty creates the dilemma of not being able to “explain” why the same level of complexity (as set by the task and the situation) results in different individuals experiencing varying levels of difficulty (as observed via differences in performance scores). The first problem creates the risk of a tautological trap that is often associated with operational definitions of constructs; the second problem seems to negate the role of the individual or person and therefore promotes a rather “un-psychological” perspective *per se*. For CPS to be taken beyond a predominantly

<sup>1</sup> We emphasize *expected* here because it highlights that complexity is based on *ex ante* developed expectations regarding the set of cognitive behaviors necessary to deal with the challenge posed by the task in conjunction with the situation.

psychometric approach, the differentiation between complexity and difficulty is a necessary precondition. Otherwise, by remaining too narrowly focussed at a psychometric level, CPS could just as appropriately be labeled Difficult Problem Solving (i.e., DPS) – a term which can be readily recognized as data driven and theoretically vacuous.

Projecting CPS-related research and its findings onto the tripartite system of Person, Task, and Situation as briefly outlined above provides a framework for the necessary differentiation between complexity and difficulty. In the following we use the semantic effect (Beckmann, 1994; Beckmann and Guthke, 1995; Beckmann and Goode, 2014), as an exemplar for how the PTS-based differentiation between complexity and difficulty can help take CPS beyond a psychometric perspective and subsequently gain a better understanding of the cognitive mechanisms behind this effect.

## VARIABLE LABELS AS SITUATIONAL CHARACTERISTIC – A SOURCE OF COMPLEXITY OR DIFFICULTY?

Previous research has repeatedly shown that seemingly minor changes of situational characteristics such as using semantically laden labels for system variables in comparison to semantically neutral labels have profound effects on performance (Beckmann, 1994; Beckmann and Guthke, 1995; Beckmann and Goode, 2014). In these studies, problem solvers tend to acquire less knowledge and subsequently control the system rather poorly when the same system is presented as a ‘Cherry Tree’ with input variables labeled ‘Light,’ ‘Water,’ and ‘Temperature’ linked to output variables labeled ‘Cherries,’ ‘Leaves,’ and ‘Beetles’ in comparison to a ‘Machine’ with input variables labeled as control dials ‘A,’ ‘B,’ ‘C’ and output variables labeled display ‘X,’ ‘Y,’ ‘Z.’ This phenomenon was initially described as the ‘Semantic Effect’ (e.g., Beckmann, 1994).

Projecting the semantic effect onto the tripartite framework of Person, Task, and Situation (PTS) implies that whilst presenting problem solvers with the same system (i.e., keeping the *task qua task* constant) and instructing them to execute the same tasks (i.e., keeping *tasks as behavior requirements* constant) still creates systematic variability in performance (i.e., indicating differences in difficulty) when a *situational characteristic* (e.g., the semantic meaning of variable labels) is varied.

As previous research has suggested, problem solvers confronted with system labels high in semanticity tend to approach the task of exploring a complex, dynamic system with a set of presumptions regarding the interrelatedness of system variables, whilst problem solvers working with variable labels low in semanticity tend to start with a “blank slate” concerning the causal structure of the system (Beckmann and Goode, 2014). In the former situation, knowledge acquisition would require a process of systematically *eliminating* presumed, yet not existing relationships and therefore reducing the complexity of the internal representation of the system’s causal structure. In the latter situation, knowledge acquisition from a “blank slate” perspective would require a process of

systematically *compiling* knowledge and therefore building up the complexity of the internal representation of the system’s causal structure. Predicting whether the cognitive processes involved in eliminating presumptions are *more complex* than those in relation to compiling knowledge would be challenging from a purely psychometric perspective.

Concomitantly, the observed performance differences in the context of the semantic effect are associated with differences in the *systematicity* of the exploration behavior (Beckmann and Goode, 2014). Systematicity in exploration behavior is reflected in a specific sequence of interventions. First, all inputs are left at zero. Any changes in the outputs can then be interpreted as autonomic changes (i.e., eigendynamics). Subsequent interventions should then focus on the effects of each input variable on any of the output variables in isolation, i.e., changing only one input at a time. Only such “Vary-One-or-None-At-a-Time” heuristic (VONAT, see Beckmann and Goode, 2014; p. 279; Beckmann and Goode, 2017) creates informative system state transitions that allow problem solvers to derive knowledge regarding the causal structure of the system. In contrast, changing more than one variable at a time or to miss the zero change intervention creates what de Jong and van Joolingen (1998, p. 185) describe as “inconclusive experiments,” which impedes successful knowledge acquisition.

## AIMS AND HYPOTHESES

We use the phenomenon of the semantic effect as an exemplary case for testing the conceptual assumption derived from the PTS framework that situational variables – in addition to task variables – present a potential source of complexity.

First, and based on findings from previous research (e.g., Beckmann and Goode, 2014) we expect problem solvers working with variable labels high in semanticity to be less systematic in their exploration behavior (*Systematicity Hypothesis*). We then address the question whether the inferior CPS performance observed under semantically rich conditions (i.e., the semantic effect) can be explained by (1) supposedly higher cognitive demands associated with a process of reducing the complexity of an internal representation of the causal structure of the explored system, or by (2) problem solvers “simply” not employing the appropriate heuristic (i.e., systematically testing against a priori assumptions). In the context of the PTS framework, results in accordance to (1) would recommend situational variables as contributors to *complexity*; results in accordance to (2) would suggest that situational variables contribute to the *difficulty* of dealing with a complex dynamic system.

The validity of a conceptual distinction between complexity and difficulty, which is based on the PTS framework, can be tested by observing the effect of explicitly instructing problem solvers to systematically explore the system by either starting with the presumption that all relationships exist (thus requiring to eliminate non-existing relationships and to reduce the complexity of the mental representation of the causal structure) or by starting with the presumption that no relationships exist (thus requiring to compile the set of relationships that exist and to build up the

complexity of the mental representation of the causal structure). If eliminating presumptions to arrive at the correct model of the system is *more complex* (via imposing greater demands on cognitive behavior) than starting with a blank slate then performance scores should worsen (both knowledge and control performance). Should, however, performance scores improve, this would suggest that problem solvers fail to engage in cognitive behavior that they are in fact capable of (*Complexity–Difficulty Hypothesis*). In psychometric terminology, the latter outcome would suggest that semanticity has the potential of introducing “construct-irrelevant difficulty” (Messick, 1995), and therefore represents a threat to validity.

## METHODS

### Participants

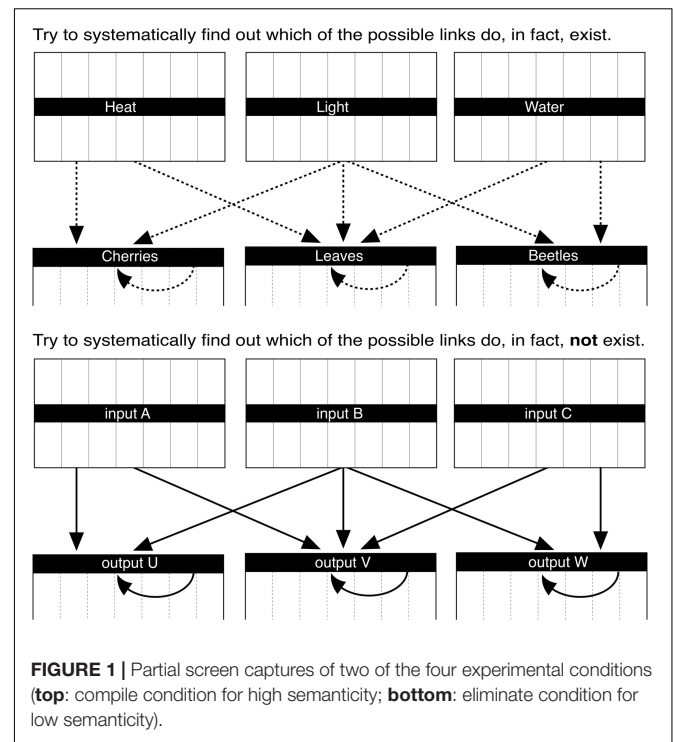
The sample comprised of 240 students from two Australian universities across a wide range of subjects, including engineering, business studies, science related subjects and medicine (60% female, mean age 22.7 years).

### Materials

To test both the Semanticity-Hypothesis and the Complexity–Difficulty Hypothesis four different versions of a CPS scenario with three input and three output variables were created (**Figure A1** shows the causal diagram and the underpinning equations that govern this system). These four versions were embedded in a 2 (semanticity: high vs. low) by 2 (instruction: compile vs. eliminate) between subject design. In the two high semanticity versions, variable labels related to a Cherry Tree were used (i.e., ‘HEAT,’ ‘LIGHT,’ and ‘WATER’ for the input variables and ‘CHERRIES,’ ‘LEAVES,’ and ‘BEETLES’ for the output variables). In the two low semanticity versions, variables low in semantic value were used (i.e., ‘INPUT A,’ ‘INPUT B,’ and ‘INPUT C’ and ‘OUTPUT U,’ ‘OUTPUT V,’ and ‘OUTPUT W,’ respectively), referring to a ‘BLACK BOX.’ For each of the two semanticity conditions two instruction conditions were created. In the compile conditions problem solvers were instructed to explore the causal structure of the given system by starting with ‘... the assumption that no relationship existed, and to systematically find out which of the possible links do, in fact, exist.’ In the eliminate conditions problem solvers were instructed to explore the causal structure of the system by starting with ‘... the assumption that all the relationships existed, and to systematically find out which of the possible links do, in fact, not exist.’

### Procedure

After completing a demographics questionnaire participants were randomly allocated to one of the four CPS conditions. The CPS systems were presented in a non-numerical, graphical format on the computer screen (see **Figure 1**). Prior to being instructed to start exploring the system under the assumption that either all or none of the relationships existed, participants allocated to the high semanticity condition (i.e., Cherry Tree) were asked to indicate their expectations regarding the causal



structure that might underpin the system. This information was used to test whether the actual implemented causal structure could be perceived as counterfactual to “common” expectations.

**Phase 1 – Knowledge Acquisition:** Participants were first instructed to acquire knowledge of the system variables’ interconnectedness. To do so they were given two cycles with seven trials each where they could freely change the values of the three input variables in their respective system and observe the subsequent changes in the output variables. After each exploration trial participants were asked to record their insights regarding the causal structure of the system in form of a causal diagram presented on screen. After the first cycle of seven trials the values for the output variables were reset, the causal diagram, however, remained on the screen.

In the compile conditions, the initial causal diagram consisted of dotted arrows representing *possible* links (see left panel in **Figure 1** for the Cherry Tree version). Over the course of the knowledge acquisition phase these arrows had to be changed into either solid arrows (indicating assumed links) or deleted arrows (indicating assumed non-links).

In the eliminate conditions, the initial causal diagram comprised solid arrows for all *possible* relationships (see right panel in **Figure 1** for the Black Box version). During the process of knowledge acquisition in this condition, arrows linking variables that were in fact identified as being unrelated were expected to be deleted from the diagram leaving only those arrows in the causal diagram for which a link is assumed to exist.

**Phase 2 – Control:** In the second phase, participants were asked to control their respective system using their developed causal diagram, which represented their previously acquired



causal knowledge. Participants had two control cycles with seven intervention trials each to reach and maintain two different target states, which were indicated as red horizontal lines in the respective panels of the output variables (Figure 2). After the first control cycle, the values for the output variables were reset and a different set of target values were given. Problem solvers were not informed about these target states prior to the respective control phase.

## Operationalizations

**Systematicity** of exploration behavior is operationalized via three ordinal categories. The intervention sequence necessary for being able to identify the underlying causal structure of the explored system comprises one exploration intervention where all inputs were left at zero followed by three exploration interventions where only one input was changed. Problem solvers who executed this sequence in this order at least once across their 14 exploration trials received a systematicity score of 2 (VONAT). Those who either failed to employ the zero intervention or where it did not precede the three single change interventions (i.e., traditional VOTAT) received a systematicity score of 1, otherwise a score of 0 was given. The rationale is that in systems with autoregressive dependencies an all preceding zero intervention is a necessary precondition to have the chance to correctly identifying direct effects using subsequent single change interventions.

**Knowledge acquisition.** The task of exploring a system to find out its underlying causal structure can be conceptualized as a “relationship detection task.” Taking CPS beyond a mere psychometric approach (i.e., by looking at more than the percentage of correctly identified relationships) should be reflected in the performance score used. We therefore based the operationalization of knowledge acquisition performance on a signal detection model that Snodgrass and Corwin (1988) introduced in the context of recognition memory. In this model the combined probability of correctly identifying existing and non-existing relationships (i.e., hits and correct rejections, resp.) form the sensitivity index  $P_r$  (Formula 1). Knowledge scores based on this operationalization have a theoretical range from  $-0.98$  to  $0.98$ , where a score below zero indicates inaccurate knowledge, whilst a score above zero indicates more accurate knowledge.

$$P_r = (\text{Hit rate}) - (\text{False Alarm rate}) \quad (1)$$

In this model a Bias Index ( $B_r$ ) can also be derived, which reflects a problem solver’s tendency to either “see” or “not to see” relationships when in fact they are uncertain. Bias scores ( $B_r$ , Formula 2) range from 0 to 1, where values below 0.5 indicate a conservative response tendency (i.e., “guessing that relationships do not exist”) and values above 0.5 indicate a liberal response tendency (i.e., “guessing that relationships exist”).

$$B_r = \frac{\text{False Alarm rate}}{1 - (\text{Hit rate}) - (\text{False Alarm rate})} \quad (2)$$

**Control performance.** An operationalization of control performance by means of a simple metric of the distance between

actual and target state after the final control intervention with limited reflection of the process, resembles the psychometric notion of a criterion-based assessment. That is, it does not differentiate between problem solvers who have reached the target state earlier and having to spend most of the time stabilizing the system, and problem solvers who reached the target closer to the end of the control cycle. As discussed in the context of measuring knowledge acquisition, given our aim to take CPS beyond a psychometric approach, the operationalization of control performance needs to better reflect how problem solvers cope with the cognitive demands (i.e., complexity) imposed by the start-target state discrepancy and the system characteristics (e.g., the dependency structure of output variables).

Finding the correct control intervention (i.e., set of inputs) that brings the system at or closest to the target state can be conceptualized as navigating the problem space. Different systems differ in their size and navigability as a function of (a) system characteristics such as the number and kind of dependencies, and/or (b) situational characteristics, such as the start-target discrepancy problem solvers must bridge. In order to allow for comparisons of performance scores across different studies, using different systems and/or different start-target discrepancies, performance scores need to be standardized against the size of the problem space of the respective system and start-target discrepancy. To achieve this standardization, we propose to operationalize control performance via the Euclidean distance between the intervention vector (i.e., values entered for the input variables) used by the participant and the vector of optimal interventions (i.e., inputs that would have brought the outputs at or closest to the target states<sup>2</sup>) for each trial of a control cycle (i.e., at each decision-input point). A standardization against the size of the problem space can be achieved by dividing the trial specific deviation scores by the trial specific difference between the vectors of pessimal and optimal intervention inputs [see formula (3)]. Consequently, control performance scores represent the averaged (across the seven control trials) deviation of the actual from the optimal intervention relative to maximal possible deviation for each and every trial. Their theoretical range is from 0 (worst possible, i.e., pessimal) to 1 (i.e., optimal).

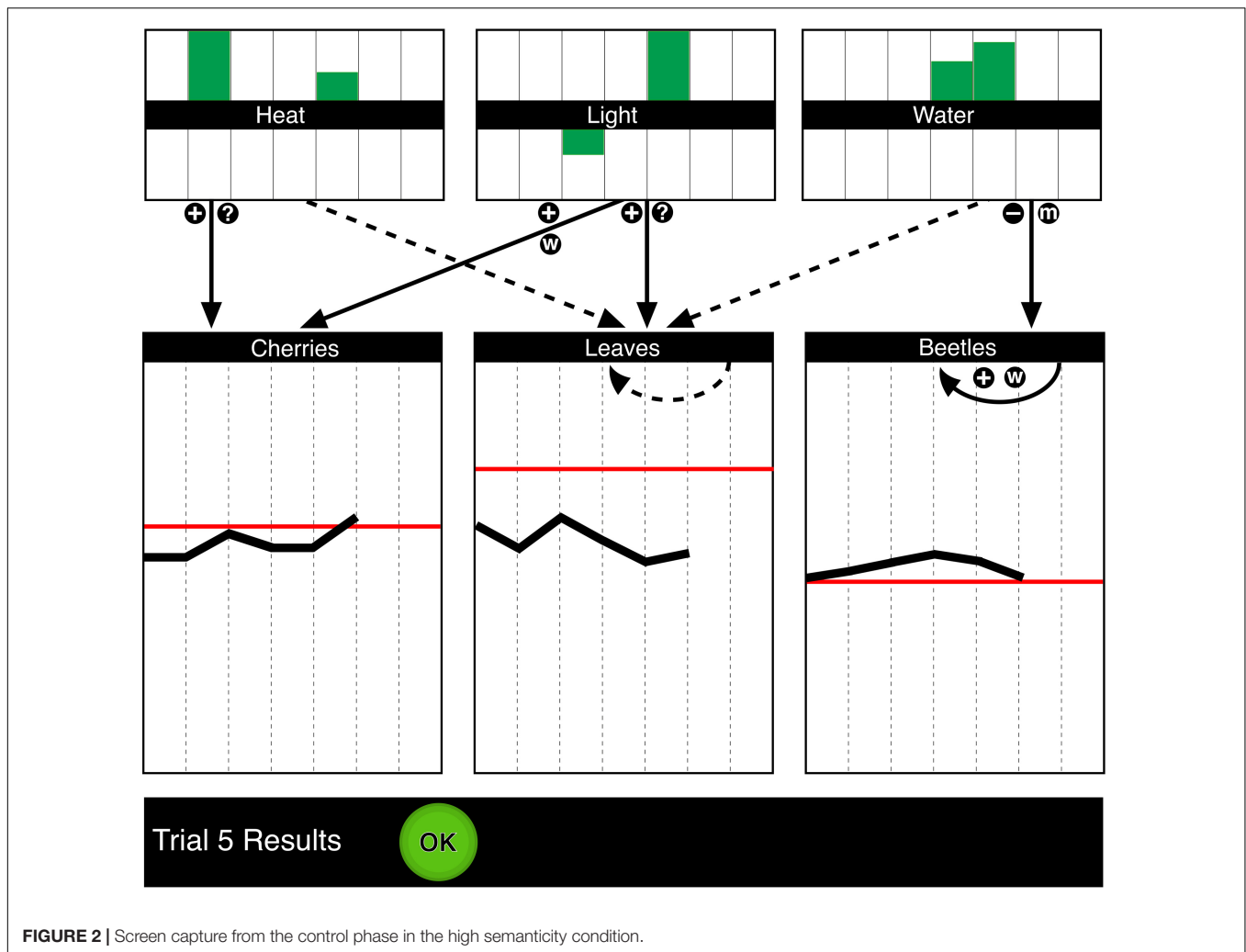
$$avEuXr = \frac{1}{m} \sum_{t=1}^m \left\{ 1 - \left[ \frac{\sqrt{\sum_{i=1}^k (\text{optimal}_{ti} - \text{actual}_{ti})^2}}{\sqrt{\sum_{i=1}^k (\text{pessimal}_{ti} - \text{optimal}_{ti})^2}} \right] \right\} \quad (3)$$

$m$ : number of trials across control cycles (14 in this study),

$k$ : number of input variables (three in this study).

<sup>2</sup>The vector of ideal inputs would bring the system exactly to (or would maintain) the target state. Restrictions of the range of possible input variables (e.g., introduced for practical reasons) might prevent reaching the target state in any one single intervention. In such cases (i.e., the ideal values fall outside this range), the values were adjusted to the nearest possible values, and these then constituted the vector of optimal inputs. In cases when the ideal values are within the range of possible inputs, the ideal values were used as the optimal input.





**FIGURE 2 |** Screen capture from the control phase in the high semanticity condition.

## RESULTS

The analyses are presented in two parts. First, we test two prerequisites, (1) the potential incompatibility of the underlying causal structure with the common expectation associated with a “real” cherry tree, and (2) the effectiveness of the instructions to start with the assumption that either all relationships existed or none of the relationships existed (manipulation check). In the second set of analyses we focus on the *Systematicity-Hypothesis* and the *Complexity-Difficulty Hypothesis*. **Table 1** provides an overview of the descriptive statistics in study-related variables across the experimental groups.

As a first step, we tested whether a potential semanticity effect might simply be explained by the causal structure that underpins the CPS system being counterfactual to what one would expect in a “real” cherry tree. To this end we analyzed the problem solvers’ expectations regarding the causal structure of the Cherry Tree prior to being instructed to explore the system (i.e., using the Sensitivity Index  $P_r$  to operationalize prior expectations as prior knowledge). The resulting average  $P_r(0)$  of  $-0.03$  ( $SD = 0.21$  based on  $N_{CT} = 124$ ) indicates

no systematic misalignment of common expectations with the actual causal structure (see **Figure A1**). In the case where the implemented system structure stood in contrast to common expectations (i.e., being “counterfactual”) the sensitivity index would have been substantially closer to  $-1.00$ . In cases where the implemented system structure would agree with a commonly held set of expectations – if such consensus existed in the first place – the resulting sensitivity index would be closer to  $+1.00$ . In the latter case, problem solvers would have already possessed knowledge that they were expected to acquire during the subsequent exploration phase. Both the average  $P_r$  value of around zero and the fact that expectations regarding the existence of relationships are equally distributed across the 12 possible variable links replicates what was found in earlier studies contrasting CPS scenarios with high and low semanticity (Beckmann, 1994; Beckmann and Goode, 2014, 2017). A counterfactual causal structure can therefore be ruled out as an alternative explanation for a potential semantic effect.

**Instruction Manipulation:** In a next step, we checked whether the instruction to start with the assumption that either all relationships existed (eliminate condition) or none

**TABLE 1 |** Descriptive statistics.

Conditions			Variables			
Semanticity	Instruction	N	Bias $B_r[1]$ M (SD)	Systematicity [No,VOTAT,VONAT] frequencies	Knowledge acquisition $P_r[14]$ M (SD)	System control $avEuXr$ M (SD)
Low (Black Box)	Compile	57	0.17 (0.19)	[3,28,26]	0.57 (0.30)	0.65 (0.12)
	Eliminate	59	0.89 (0.19)	[15,25,19]	0.41 (0.35)	0.60 (0.12)
High (Cherry Tree)	Compile	63	0.35 (0.27)	[18,33,12]	0.20 (0.25)	0.55 (0.09)
	Eliminate	61	0.87 (0.20)	[14,34,13]	0.33 (0.36)	0.59 (0.13)

of the relationships existed (compile condition) was reflected in problem solvers' response behavior during the knowledge acquisition phase. Problem solvers' ability to follow these instructions should be identifiable in the trajectories of the bias scores ( $B_r$ ) over the course of the two exploration cycles with their seven trials each. We expect problem solvers in the eliminate condition to start with a bias score greater than 0.5 and close to 1.00 as this would indicate an instruction-induced tendency "to guess that there is" a relationship when in fact (still) in a state of not knowing. Problem solvers in the compile condition, however, were expected to start with a conservative bias (i.e., a  $B_r$  score below 0.5 and close to zero), which would indicate a response tendency of "guessing that there is not" a relationship when in the state of (yet) not knowing. In both conditions, we expected bias scores to become more neutral (i.e.,  $B_r \approx 0.5$ ) over the course of the exploration trials and when progressing in acquiring knowledge. The left panel in **Figure 3** depicts the differing bias trajectories for the "Black Box" conditions; the right panel shows them for the "Cherry Tree" conditions. It is interesting to note that the final convergence occurs at a level of around 0.75 for all conditions. This seems to indicate a general propensity to slightly err on the positive, i.e., to rather assume that there are relationships than running the risk of missing one.

The trajectories seem to suggest that the instruction has led to the expected differences in response behavior, confirming the effectiveness of the instruction manipulation, in general. Two further suggestions seem to emerge. First, the slopes for the eliminate conditions are markedly less steep than the ones for the compile conditions ( $F_{4,1,968.5} = 76.455$ ,  $p < 0.001$ ,  $\eta^2 = 0.224$ )<sup>3</sup>, which seems to suggest that reducing complexity is more challenging than increasing it, regardless of semanticity. Second, the starting point of the compile condition for "Cherry Tree" is not as low as it is for "Black Box" ( $B_r[1]$ :  $t_{118} = 4.13$ ,  $p < 0.001$ ,  $d_{\text{compile-BBvsCT}} = 0.76$ ), which seems to suggest that adopting a "blank slate" perspective is more challenging in a system with high semanticity.

## Systematicity

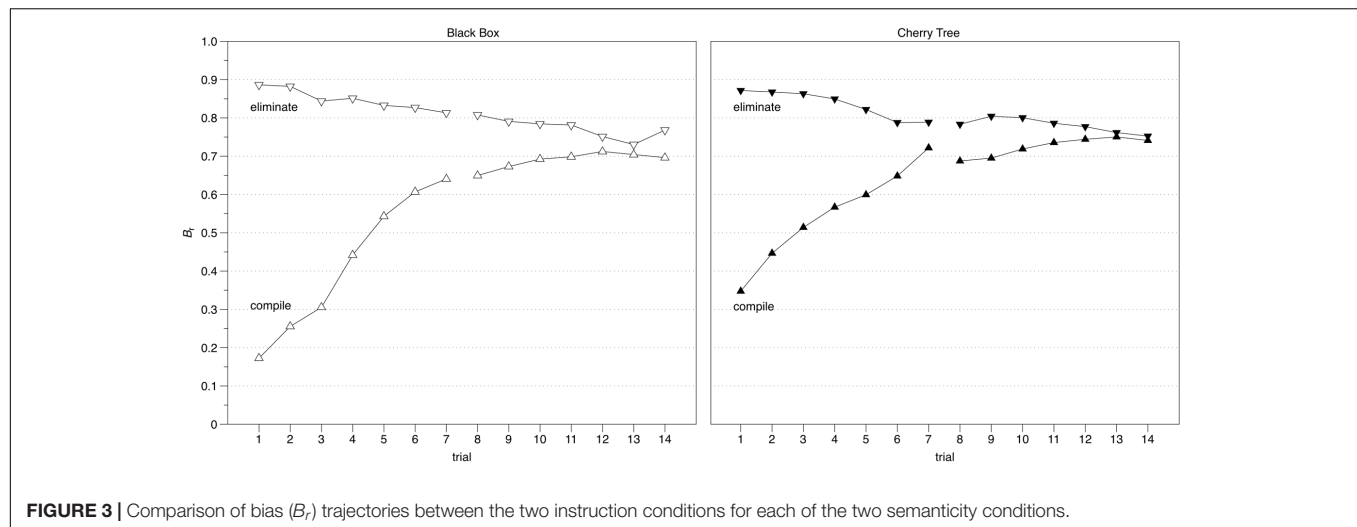
To test the Systematicity-Hypothesis we conducted an ordinal logistic regression analysis where problem solvers' VONAT score

was regressed on the semanticity condition and the instruction condition they have been allocated to. The results indicate (see **Table 2**) that problem solvers who were asked to explore a system with low semanticity (i.e., Black Box) were 2.24 time more likely to employ a systematic exploration heuristic (i.e., using VOTAT or VONAT) than problem solvers working on a system that used variables labels high in semanticity (i.e., Cherry Tree). Being instructed to either systematically eliminate erroneously presumed relationships or to identify existing relationships in the causal model of the respective system did not, however, make a substantial difference in the level of systematicity with which problem solvers explored the system.

## Complexity – Difficulty

To address the *Complexity-Difficulty Hypothesis* we tested in a final step whether the effect of the instruction differs between the two levels of semanticity in terms of the knowledge acquisition performance and control performance metrics. Given both metrics produced comparable effects and interpretations, we report them together. As expected, the ANOVAs resulted in a main effect of the situational factor "semanticity," with overall lower performance scores (knowledge acquisition:  $F_{1,236} = 29.863$ ,  $p < 0.001$ ,  $\eta^2 = 0.11$ ; control performance:  $F_{1,236} = 14.048$ ,  $p < 0.001$ ,  $\eta^2 = 0.06$ ) for the high semanticity condition (i.e., Cherry Tree) relative to the low semanticity condition ("Black Box"). This replicates the semantic effect once more (Beckmann, 1994; Beckmann and Guthke, 1995; Beckmann and Goode, 2014). Across the two semanticity conditions, the task factor "instruction" seems to have no effect on performance scores overall (knowledge acquisition:  $F_{1,236} = 0.119$ ,  $p = 0.730$ ,  $\eta^2 \approx 0.00$ , control performance:  $F_{1,236} = 0.027$ ,  $p = 0.870$ ,  $\eta^2 \approx 0.00$ ). However, the presence of an interaction effect (knowledge acquisition:  $F_{1,236} = 13.235$ ,  $p < 0.001$ ,  $\eta^2 = 0.05$ , control performance:  $F_{1,236} = 7.544$ ,  $p = 0.006$ ,  $\eta^2 = 0.03$ ), indicates that when being instructed to start with the assumption that all relationships existed and consequently systematically eliminate unjustified presumptions showed a positive effect on both knowledge acquisition and control performance in conditions of high semanticity (i.e., Cherry Tree), but it resulted in systematically lower performance scores in the condition where problem solvers were working with low levels of semanticity (i.e., Black Box, see **Figure 4**).

<sup>3</sup>Greenhouse-Geisser correction of  $dfs$  for the  $F$ -test was applied due to sphericity ( $\chi^2_{90} = 2201.984$ ,  $p < 0.001$ ).



**TABLE 2 |** Ordinal logistic regression of systematicity (VONAT) on semanticity and instruction.

	Estimate	SE	Wald $\chi^2$	df	p	Odds ratio
Semanticity (Black Box vs. Cherry Tree)	0.804	0.251	10.256	1	0.001	2.24
Instruction (compile vs. eliminate)	−0.307	0.246	1.557	1	0.212	0.74

## Summary

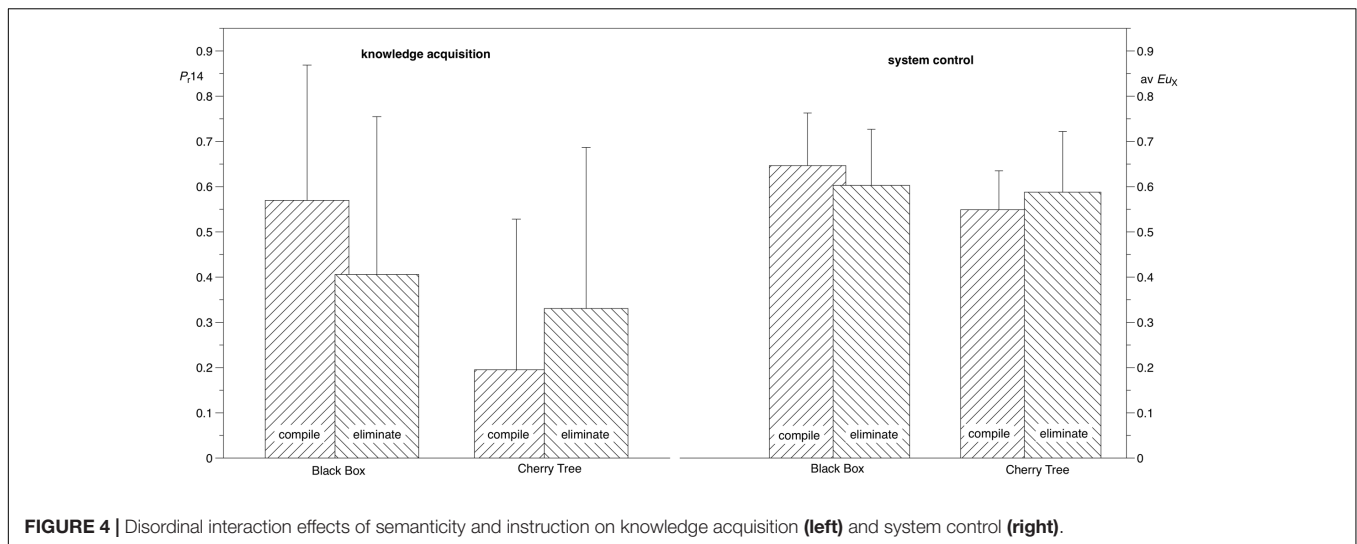
Knowledge acquisition, especially in systems with high levels of semanticity, can be conceptualized as a process of transforming a presumption structure into a knowledge structure. The instruction to start with the assumption that all possible relationships between system variables existed aimed at creating a presumption structure with high levels of complexity. If we were to use the number of relationships in a system ( $NoR$ ) as a crude quantifier of complexity (see Beckmann and Goode, 2017 for a more detailed discussion) the process of knowledge acquisition under this instruction requires the reduction of complexity from a  $NoR_{presumed} = 12$  to  $NoR_{actual} = 6$ . In contrast, the instruction to start with a “blank slate” (i.e., assuming that no relationship exists) aimed at creating a situation where complexity needed to be increased from  $NoR_{presumed} = 0$  to  $NoR_{actual} = 6$ . The slope differences in bias scores between the two instruction conditions suggest that decreasing the complexity of a presumption structure is more challenging than is building up the complexity of a knowledge structure, regardless of the semanticity of the explored system.

If we were to interpret the difference in the initial Bias-scores between the two instruction conditions as an indicator of how well problem solvers were able to adopt a “full slate” or “blank slate” perspective then the significant interaction effect between semanticity and instruction would indicate that the *instruction-adoption* differs between the two semanticity conditions. Problem solvers tend to struggle adopting a “blank slate” perspective under the high semanticity condition. From

a cognitive task analysis point of view, we could surmise that adopting a “blank slate” perspective under high semanticity conditions requires the suppression of preconceived expectations regarding the causal structure of the system as they are triggered by the semanticity of the variable labels. The process of suppression or decontextualization seems to add to the *complexity* of the task of knowledge acquisition in CPS-systems high in semanticity. In short, *semanticity*, as a situational characteristic of CPS, has the potential of being a source of *complexity*.

We have also argued that systematicity (i.e., the creation of informative mini-experiments that help to identify the existence or non-existence of relationships between system variables) is a necessary precondition for successful knowledge acquisition, independent of instruction conditions or semantic embedding of the system. Our findings, however, suggest that problem solvers working under high semanticity conditions are on average less likely to engage in systematic exploration behavior. At this stage, it is difficult to conceive of a “cognitive argument” that would predict that the heuristic of systematically *testing against* presumptions (as required in the eliminate conditions) is cognitively more demanding than *testing for* evidence of the existence of relationships (as would be required in the compile conditions). This, in conjunction with the fact that problem solvers in the compile condition with low semanticity were able to be more systematic, leads to the conjecture of seeing the failing to employ a suitable or necessary heuristic as an indication of the greater *difficulties* problem solvers seem to have. In short, semanticity as a situational characteristic of CPS might also be a potential source of (unnecessary) *difficulty*.

In switching the focus from the bias score (indicating the adoption of the instructed behavior) and systematicity score (indicating the level of engaging in planned and coordinated behavior) onto performance (i.e., knowledge acquisition as well as control), the data suggest that in conditions of high semanticity, it is more effective to start with the presumption that all relationships might exist (“full slate”) rather than to start pretending that none exist. This requires systematic *testing*



**FIGURE 4 |** Disordinal interaction effects of semanticity and instruction on knowledge acquisition (left) and system control (right).

against a priori assumptions regarding the system's underlying causal structure, with the emphasis on "systematic." In conditions with low semanticity, however, it seems *less* effective to start with presumptions of existing relationships (it is safe to assume that such presumptions would likely be a result of conscious efforts to guess). The more "natural" starting position here would be something more akin to a "blank slate" (or, knowing that one does not know), which then would require a systematic *testing* for evidence regarding the system's underlying causal structure.

The fact that knowledge acquisition and control performances in the high semanticity conditions still fall short of those shown in low semanticity conditions (i.e., replicating the "semantic effect") can be explained via two factors. First a *complexity factor*, which reflects the additional cognitive demands associated with suppressing presumptions when trying to adopt a "blank slate" starting position under high semanticity conditions, and second a *difficulty factor*, which reflects the tendency of problem solvers to not adopt a systematic approach to exploration behavior.

## GENERAL DISCUSSION

These reflections should not be misunderstood as an unconditional plea against the use of semantically laden variable labels in CPS. The answer to the question of what kind of systems should be used is once more the infamous: it depends. It depends on the purpose of the use of CPS scenarios. If, for instance, we aim to measure problem solvers' ability to draw inferences based on observed outcomes of systematic experimentation, we need to consider that arguably minor changes in situational characteristics, such as the semanticity of variable labels, have the potential to prevent the spontaneous employment of systematic experimentation (see also Beckmann and Goode, 2014). Under these circumstances, it would be inappropriate to interpret performance scores as indicators of problem solvers' reasoning ability or to expect them to correlate highly with reasoning measures. If, however, the aim was to predict "real life decision making" and given that "real life

problems" are always semantically anchored, then using systems with high semanticity might be appropriate. The "construct purity" (or uni-dimensionality, in psychometric terms) of the measure, however, is likely to be compromised, which needs to be reflected (a) in expectations regarding inter-test correlations and (b) in the way performance scores are interpreted. Schoppek and Fischer (2015) make a convincing case for conceptualizing CPS performance scores as indicators of a competency, whereby a competency is a conglomerate of knowledge, reasoning ability, thinking skills and motivational variables. The PTS framework proposed here can help draw attention to the often-overlooked potential impact that situational characteristics might have on the composition of knowledge, reasoning ability, thinking skills and motivational variables in performance scores obtained from dealing with supposedly homomorphous CPS systems. For instance, the practice of aggregating performance scores obtained in multiple minimal complex systems (e.g., Funke, 2014; Stadler et al., 2016) with various levels of semanticity might be psychometrically desirable (e.g., maximizing reliability). At the same time, however, this very practice could (inadvertently) turn out to be a threat to construct validity if performance scores are underpinned by qualitatively different cognitive processes (e.g., compiling vs. eliminating), varying levels of functional or dysfunctional prior knowledge, and/or differences in perceived personal relevance of the semantic context these systems are embedded in. The PTS framework might also be (in-)formative for on-going discussions as to whether CPS performance scores are more than *g* or not (e.g., Gonzales et al., 2005; Wüstenberg et al., 2012; Hundertmark et al., 2015; Schoppek and Fischer, 2015; Stadler et al., 2015).

The use of semantically laden cover stories or variable labels to induce a stronger sense of "real life" relevance of the CPS experience for the participants in our laboratory studies or large scale assessment exercises should also not be mistaken as a shortcut to what some might call ecological validity. If we were to define ecological validity as the meaningfulness, appropriateness and usefulness of inferences drawn based on



performance scores obtained using an assessment tool, one would have to convincingly demonstrate that the problems posed in the assessment situation have triggered the same cognitive or affective processes as they are expected to be involved when dealing with complexity, uncertainty and dynamics in the “real world.” Otherwise we run the risk of simply falling prey to our own make-believe. Proper validation requires an *ex ante* specification of the cognitive or affective processes expected to be involved when dealing with complexity, uncertainty and dynamics in the “real world.” The distinction between complexity and difficulty, as being proposed here, can help moving beyond psychometrics-driven *post hoc* interpretations of mean scores and correlation patterns.

The differentiation between complexity and difficulty can also help improving the conceptual and psychometric quality of the assessment or research tools we use in the context of CPS. For instance, result patterns indicating that problem solvers overall experienced fewer difficulties (i.e., better performance) than the *ex ante* specifications of complexity would have led us to expect, could suggest that problem solvers might not have had to engage in the sequences of cognitive processes anticipated. Certain task-independent situational features could have enabled the use of prior knowledge or chunked information cues (Wood, 1986) and consequently created “construct-irrelevant easiness” (Messick, 1995). Conversely, “construct-irrelevant difficulty” (Messick, 1995) could result from a misalignment between (empirically observed) performance scores and (theoretically pre-determined) complexity specifications, where the former is systematically lower than the latter would have suggested. This could have been triggered by situational variables (inadvertently) preventing problem solvers from engaging in the anticipated sequence of cognitive behaviors. Both instances present a threat to validity that might be overlooked if complexity and difficulty are treated synonymously.

The main intent of this paper was to contribute to the discussion around taking CPS beyond a narrowly defined psychometric approach. We are of the view that a predominantly psychometric perspective tends to fall short in appropriately capturing the essence of CPS, namely complexity. We identified the lack of a differentiation between complexity and difficulty as a major barrier to achieving conceptual progress in CPS research. To redress this, we introduced the Person–Task–Situation (PTS) framework which, through the theoretical distinction it makes between its constituent factors, enables a conceptual differentiation of complexity and difficulty. The differentiation provides a theory-based platform for studying cognition (e.g., information processing, learning, decision making, reasoning) beyond an atheoretical psychometric lens.

Complexity as a concept also includes a qualitative dimension, whilst difficulty is exclusively quantitative. Complexity is a cognitive concept that reflects the interactive effects of information processing demands imposed upon the cognitive system by task and situation characteristics (i.e., the T and the S in the PTS framework). Difficulty is a psychometric concept that reflects the level of success problem solvers have in dealing with complexity. The integration of the person (i.e., the P in the PTS framework) introduces individual differences in ability, memory,

knowledge and attitudinal variables as potential explanatory factors for observed performance differences. Cognition research in general, and CPS research in particular, focuses on studying the links between complexity and difficulty. By ignoring their conceptual differences and treating them synonymously, CPS research runs the risk of losing sight of its cognition-based origins and failing to utilize its potential.

As a case in point, we used the “semantic effect” to test these conceptualizations. We were able to show that by using the same system (i.e., keeping the *task qua task* constant) and asking problem solvers to freely explore the system to find out its underlying causal structure (i.e., keeping the *task as behavior requirement* constant), but varying the system’s semantic embedment via using different variable labels (i.e., varying a *situational* variable) systematic differences in exploration behavior occurred. Failing to differentiate task and situation as independent sources of complexity and by treating complexity and difficulty synonymously the resulting performance differences would erroneously be attributed to individual differences in person-related variables.

The conceptual distinction between complexity and difficulty paves the path for taking CPS beyond a psychometric approach. In fact, it is instrumental to bringing the “psycho-” back into psychometric. Otherwise one tends to operate with a “metric” that is agnostic to theory, and can therefore not be scrutinized for validity. The validity question is the core element of empirical research in psychology that relies on a strong conceptual underpinning. Psychometrics is a tool for linking the theoretical and the empirical and should not be used as a substitute for either.

The study presented here is not intended as a comprehensive test of the PTS framework that underpins the complexity–difficulty distinction. Instead, the paper should be considered as an invitation and orientation for future work. The theoretical analyses and empirical outcomes we report support the proposed complexity framework in demonstrating that it is both specific enough to allow for testable hypotheses, yet broad enough to allow modifications and refinements. Our work also contributes to efforts to better understand the person–task–situation tripartite. Future conceptual and empirical contributions will be necessary to further develop and refine a common framework that considers the interplay of the person, the task and the situation and has complexity at its conceptual core. This, so we have argued, is particularly pertinent to a research paradigm such as CPS that carries complexity in its label. Efforts to this end will support the better integration of research findings from existing and future studies on CPS.

## ETHICS STATEMENT

This study has been reviewed by, and received ethics clearance through, the Human Research Ethics Committee (HREC), University of New South Wales, Australia (Approval No: 09616 and 06294). After being informed (a) what participation in the study entails, (b) that participation was voluntary and (c) withdrawal from participation was possible at any time without negative consequences, and (d) that anonymity was guaranteed,

participants were asked to sign an informed consent form prior to participation.

## AUTHOR CONTRIBUTIONS

JB, DB, and NG certify that they have participated sufficiently in the work to take responsibility for the content, including participation in the conception, design, analysis, drafting the work, writing, and final approval of the manuscript. Each author agrees to be accountable for all aspects of the work.

## REFERENCES

- Beckmann, J. F. (1994). *Lernen und Komplexes Problemlösen. [Learning and Complex Problem Solving]*. Bonn: Holos.
- Beckmann, J. F. (2010). Taming a beast of burden – on some issues with the conceptualisation and operationalisation of cognitive load. *Learn. Instr.* 20, 250–264. doi: 10.1016/j.learninstruc.2009.02.024
- Beckmann, J. F., and Goode, N. (2014). The benefit of being naïve and knowing it: the unfavourable impact of perceived context familiarity on learning in complex problem solving tasks. *Instr. Sci.* 42, 271–290. doi: 10.1007/s11251-013-9280-7
- Beckmann, J. F., and Goode, N. (2017). Missing the wood for the wrong trees: on the difficulty of defining the complexity of complex problem solving scenarios. *J. Intell.* 5:15. doi: 10.3390/jintelligence5020015
- Beckmann, J. F., and Guthke, J. (1995). “Complex problem solving, intelligence, and learning ability,” in *Complex Problem Solving. The European Perspective*, eds P. A. Frensch and J. Funke (Hillsdale, NJ: Erlbaum), 177–200.
- Birney, D. P., Beckmann, J. F., and Seah, Y.-Z. (2016). More than the eye of the beholder: the interplay of person, task, and situation factors in evaluative judgements of creativity. *Learn. Individ. Differ.* 51, 400–408. doi: 10.1016/j.lindif.2015.07.007
- de Jong, T., and van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Rev. Educ. Res.* 68, 179–201. doi: 10.3102/00346543068002179
- Frensch, P. A., and Funke, J. (eds) (1995). *Complex Problem Solving: The European Perspective*. Hillsdale, NJ: Erlbaum.
- Funke, J. (1985). Steuerung dynamischer systeme durch aufbau und anwendung subjektiver Kausalmodelle [Control of dynamic systems by building up and using subjective causal models]. *Z. Psychol.* 193, 435–457.
- Funke, J. (1992). Dealing with dynamic systems: research strategy, diagnostic approach and experimental results. *German J. Psychol.* 16, 24–43.
- Funke, J. (2014). Analysis of minimal complex systems and complex problem solving require different forms of causal cognition. *Front. Psychol.* 5:739. doi: 10.3389/fpsyg.2014.00739
- Funke, J., Fischer, A., and Holt, D. (2017). When less is less: solving multiple simple problems is not complex problem solving—A comment on Greiff et al. (2015). *J. Intell.* 5, 1–11. doi: 10.1111/desc.12216
- Funke, J., and Frensch, P. A. (2007). “Complex problem solving: the European perspective — 10 years after,” in *Learning to Solve Complex Scientific Problems*, ed. D. H. Jonassen (New York, NY: Lawrence Erlbaum), 22–47.
- Gonzales, C., Thomas, R. P., and Vanyukov, P. (2005). The relationships between cognitive ability and dynamic decision making. *Intelligence* 33, 169–186. doi: 10.1016/j.cub.2016.09.015
- Greiff, S., Stadler, M., Sonleitner, P., Wolff, C., and Martin, R. (2015). Sometimes less is more: comparing the validity of complex problem solving measures. *Intelligence* 50, 100–113. doi: 10.1016/j.intell.2015.02.007
- Guthke, J., Beckmann, J. F., and Stein, H. (1995). “Recent research evidence on the validity of learning tests,” in *Advances in Cognition and Educational Practice*.

## FUNDING

This research was supported under Australian Research Council's Linkage Projects funding scheme (project LP0669552). The views expressed herein are those of the authors and are not necessarily those of the Australian Research Council.

## ACKNOWLEDGMENT

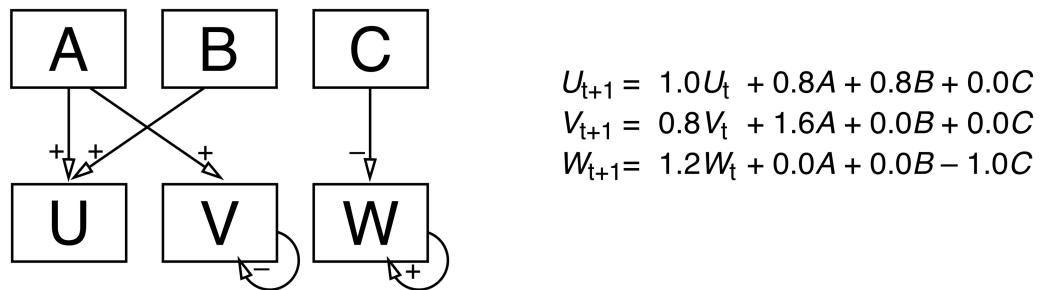
We would like to thank Anissa Müller, Myvan Bui, and Lisa Reinecke for their support with the data collection.

- European Contributions to the Dynamic Assessment*, Vol. 3, ed. J. S. Carlson (Greenwich, CT: JAI Press), 117–143.
- Hackman, J. R. (1969). Towards understanding the role of tasks in behavioural research. *Acta Psychol.* 31, 97–128. doi: 10.1016/0001-6918(69)90073-0
- Hundertmark, J., Holt, D. V., Fischer, A., Said, N., and Fischer, H. (2015). System structure and cognitive ability as predictors of performance in dynamic system control tasks. *J. Dyn. Decis. Mak.* 1, 1–10. doi: 10.11588/jddm.2015.1.26416
- McGrath, J. E., and Altman, I. (1966). *Small Group Research: A Synthesis and Critique of the Field*. New York: Holt, Rinehart, & Winston.
- Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* 50, 741–749. doi: 10.1037/0003-066X.50.9.741
- Newell, A., and Simon, H. A. (1972). *Human Information Processing*. Englewood Cliffs, NJ: Prentice-Hall.
- OECD (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. Paris: OECD Publishing.
- Osman, M. (2010). Controlling uncertainty: a review of human behavior in complex dynamic environments. *Psychol. Bull.* 136, 65–86. doi: 10.1037/a0017815
- Schoppek, W., and Fischer, A. (2015). Complex problem solving – single ability or complex phenomenon? *Front. Psychol.* 6:1669. doi: 10.3389/fpsyg.2015.01669
- Snodgrass, J. G., and Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *J. Exp. Psychol. Gen.* 117, 34–50. doi: 10.1037/0096-3445.117.1.34
- Stadler, M., Becker, N., Gödker, M., Leutner, D., and Greiff, S. (2015). Complex problem solving and intelligence: a meta-analysis. *Intelligence* 53, 92–101. doi: 10.1016/j.intell.2015.09.005
- Stadler, M., Niepel, C., and Greiff, S. (2016). Easily too difficult: estimating item difficulty in computer simulated microworlds. *Comput. Hum. Behav.* 65, 100–106. doi: 10.1016/j.chb.2016.08.025
- Wood, R. E. (1986). Task complexity: definition of the construct. *Organ. Behav. Hum. Decis. Process.* 37, 60–82. doi: 10.1016/0749-5978(86)90044-0
- Wood, R. E., Beckmann, J. F., and Birney, D. P. (2009). Simulations, learning and real world capabilities. *Educ. Train.* 51, 491–510. doi: 10.1108/00400910910987273
- Wüstenberg, S., Greiff, S., and Funke, J. (2012). Complex problem solving — More than reasoning? *Intelligence* 40, 1–14. doi: 10.1016/j.intell.2011.11.003

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Beckmann, Birney and Goode. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

## APPENDIX



**FIGURE A1** | Diagram and set of equations for the causal structure of the CPS system used in this study.



# Common Process Demands of Two Complex Dynamic Control Tasks: Transfer Is Mediated by Comprehensive Strategies

Wolfgang Schoppek<sup>1\*</sup> and Andreas Fischer<sup>2</sup>

<sup>1</sup> University of Bayreuth, Bayreuth, Germany, <sup>2</sup> Forschungsinstitut Betriebliche Bildung, Nürnberg, Germany

## OPEN ACCESS

### Edited by:

David Peebles,  
University of Huddersfield,  
United Kingdom

### Reviewed by:

C. Dominik Güss,  
University of North Florida,  
United States  
Christin Lotz,  
Saarland University, Germany

### \*Correspondence:

Wolfgang Schoppek  
wolfgang.schoppek@uni-bayreuth.de

### Specialty section:

This article was submitted to  
Cognitive Science,  
a section of the journal  
Frontiers in Psychology

**Received:** 15 March 2017

**Accepted:** 27 November 2017

**Published:** 19 December 2017

### Citation:

Schoppek W and Fischer A (2017)  
Common Process Demands of Two  
Complex Dynamic Control Tasks:  
Transfer Is Mediated by  
Comprehensive Strategies.  
Front. Psychol. 8:2145.  
doi: 10.3389/fpsyg.2017.02145

Although individual differences in complex problem solving (CPS) are well-established, relatively little is known about the process demands that are common to different dynamic control (CDC) tasks. A prominent example is the VOTAT strategy that describes the separate variation of input variables (“Vary One Thing At a Time”) for analyzing the causal structure of a system. To investigate such comprehensive knowledge elements and strategies, we devised the real-time driven CDC environment Dynamis2 and compared it with the widely used CPS test MicroDYN in a transfer experiment. One hundred sixty five subjects participated in the experiment, which completely combined the role of MicroDYN and Dynamis2 as source or target problem. Figural reasoning was assessed using a variant of the Raven Test. We found the expected substantial correlations among figural reasoning and performance in both CDC tasks. Moreover, MicroDYN and Dynamis2 share 15.4% unique variance controlling for figural reasoning. We found positive transfer from MicroDYN to Dynamis2, but no transfer in the opposite direction. Contrary to our expectation, transfer was not mediated by VOTAT but by an approach that is characterized by setting all input variables to zero after an intervention and waiting a certain time. This strategy (called PULSE strategy) enables the problem solver to observe the eigendynamics of the system. We conclude that for the study of complex problem solving it is important to employ a range of different CDC tasks in order to identify components of CPS. We propose that besides VOTAT and PULSE other comprehensive knowledge elements and strategies, which contribute to successful CPS, should be investigated. The positive transfer from MicroDYN to the more complex and dynamic Dynamis2 suggests an application of MicroDYN as training device.

**Keywords:** complex problem solving, complex dynamic control, dynamic decision making, strategies, knowledge acquisition

## INTRODUCTION

Complex problem solving (CPS) is a phenomenon that is investigated in many domains, ranging from scientific discovery learning over industrial process control to decision making in dynamic economical environments. At the heart of the scientific investigation of the phenomenon are complex dynamic control (CDC) tasks (Osman, 2010) that are simulated in the laboratory. Simulated CDC tasks provide the opportunity to study human deciding and acting in complex situations under controlled and safe conditions.



Currently, research on CPS is dominated by attempts to construe it as one-dimensional ability construct, which means that a single measure represents a person's ability to solve complex problems. To this end, Greiff and Funke (2010) and Greiff et al. (2012) have developed the minimal complex systems test MicroDYN. This CPS environment consists of a number of linear systems with mostly three input and three output variables. The systems are presented with various cover stories (e.g., how do different training schedules affect aspects of handball performance?). The subjects have to explore each system, enter their insights into a causal diagram (knowledge acquisition phase) and subsequently steer the system to a given array of target values by entering input values (knowledge application phase). Each system is attended to for about 5 min. MicroDYN yields reliable measures of knowledge acquisition and knowledge application (Fischer et al., 2015a). As both variables are highly correlated, they are often combined to obtain a measure of CPS ability (e.g., Greiff and Fischer, 2013). MicroDYN has been validated using various criteria—predominantly school grades. The typical result of these studies is that the combined CPS measure accounts for 5% variance in school grades incremental to figural reasoning (Schoppek and Fischer, 2015).

Consistent with our view of CPS as a multifaceted phenomenon (Schoppek and Fischer, 2015), we claim to use the denomination “complex problem solving” in a broader sense. We adhere to the conception of Dörner (1997), who characterizes complex problems as being complex (many variables), interrelated (with many relations among the variables), dynamic (with autonomous state changes), intransparent (with not all information being available at the outset), and polytelic (more than one goal has to be considered; often goals are contradicting). As these characteristics are not defined precisely, and can take shape to varying degrees, CPS refers to a broad range of problems, which can differ considerably in their requirements for being solved (Fischer and Neubert, 2015). This could be considered a conceptual weakness. However, for the labeling of broad phenomena this is common practice. For example, the established label “problem solving” has an even larger domain. Therefore, assuming a one-dimensional construct “CPS” does not do justice to the heterogeneity of the domain (Fischer and Neubert, 2015).

In order to make progress toward a deeper understanding of CPS we propose a preliminary process model (see **Figure 1**). The model is composed of assumptions that are established in the CPS literature. We classify these assumptions as pertaining to processes and structures.

One coarse *process* assumption divides CPS in the phases (or sub-processes) of *knowledge acquisition* and *knowledge application* (Fischer et al., 2012). Knowledge acquisition refers to the requirement of detecting the causal structure of the system by means of appropriate exploration strategies<sup>1</sup>. Knowledge

application means using the acquired knowledge to plan and implement interventions in order to reach given target states. This assumption of Fischer et al. (2012) originates in the Dynamis approach by Funke (1991, 1993) and underlies the MicroDYN paradigm (Greiff and Funke, 2009). In view of the widely spread use of this model, we call it the “standard model of CPS.” A second classification, proposed by Osman (2008, 2010), distinguishes between monitoring, which “refers to online awareness and self-evaluation of one's goal-directed actions” (Osman, 2008, p. 97), and control, which refers to “the generation and selection of goal-directed actions” (ibid., p. 97). As Osman (2008) operationalized monitoring through observation of exploration behavior of oneself or others, the kinship between monitoring and knowledge acquisition becomes obvious. However, control pertains to knowledge application and exploratory manipulations (which are part of the knowledge acquisition sub-process of the standard model).

With respect to *structure*, Schoppek (2002) has proposed a classification of knowledge types that are learned during and/or applied to CPS: *Structural knowledge* is knowledge about the causal relations among the variables that constitute a dynamic system. *I-O knowledge* (shorthand for “input-output knowledge”) represents instances of interventions together with the system's responses. *Strategy knowledge* represents abstract plans of how to cope with the CDC problem. An example is the awareness of the control of variables strategy (Chen and Klahr, 1999), also known as VOTAT (Vary One Thing At a Time, Tschirgi, 1980).

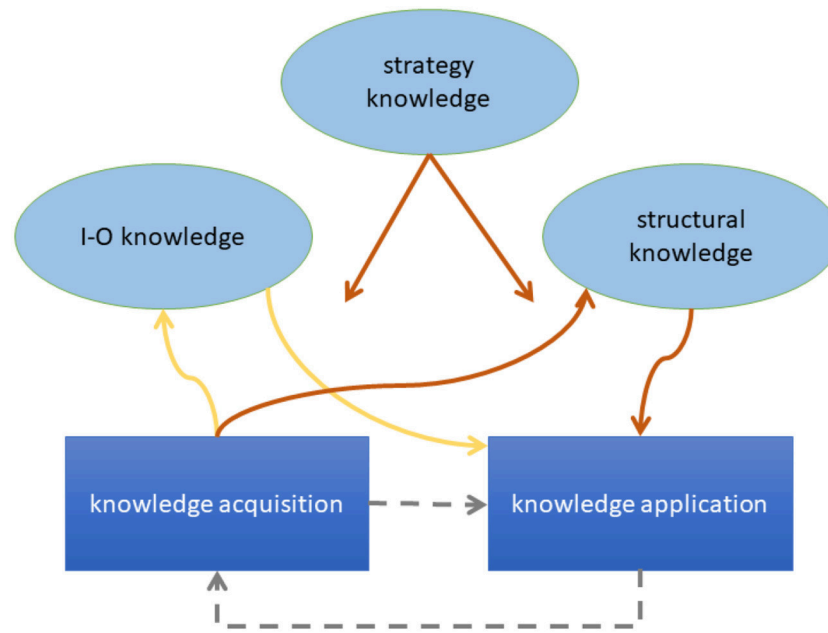
VOTAT was first described in the context of testing hypotheses in multivariate stories (Tschirgi, 1980). In the context of CDC tasks, it means varying a single input variable in order to observe its effects on the output variables. The extent of using this strategy predicts better structural knowledge and better control performance (Vollmeyer et al., 1996; Wüstenberg et al., 2014).

A related strategy is to apply an impulse to an input variable: The problem solver sets one or more input variables to certain values greater than zero, then sets the values back to zero again. In the following simulation steps where all input variables are zero, the course of the output variables informs the problem solver about side effects and eigendynamics of the output variables<sup>2</sup>. Schoppek (2002) instructed this strategy to participants in an experiment that involved a CDC task of the Dynamis type and found better structural knowledge in the trained group (see also Beckmann, 1994 and Schoppek and Fischer, 2015). Evidence about the usefulness of this strategy for controlling MicroDYN has recently been reported by Greiff et al. (2016) and Lotz et al. (2017). These authors refer to the strategy as non-interfering observation or NOTAT. We use the label PULSE, following Schoppek's (2002) characterization as setting an impulse.

Back to the process model: Processual and structural assumptions are different perspectives rather than alternative conceptions. For example, in the knowledge acquisition phase the

<sup>1</sup> Although we would prefer to distinguish between tactics (= concrete methods for accomplishing goals), and strategies (= abstract plans), we use the more general term “strategy” for both, because the distinction has not become widely accepted in cognitive science.

<sup>2</sup> A side effect is an effect of one output variable on another; eigendynamic is the effect of an output variable on itself (Funke, 1992). Considerations about how to deal with eigendynamic can be traced back to the early days of CPS (Dörner, 1980; Beckmann, 1994; Dörner and Schaub, 1994).



**FIGURE 1 |** Visualization of the preliminary process model. Brown arrows denote processes that require much working memory capacity; yellow arrows denote processes that require little working memory capacity. The arrows originating from “strategy knowledge” indicate that this knowledge determines most of the displayed processes. Gray arrows indicate that the processes can iterate within the same problem.

goal is to gain structural knowledge about a system by application of appropriate strategies such as VOTAT, which are part of the strategy knowledge of the problem solver. The execution of VOTAT in turn is a process.

Our process model includes assumptions about the transfer distance of the different knowledge types (Schoppek, 2002). Structural knowledge about one specific System A cannot be transferred to another System B with a different structure (far transfer, see Paas, 1992). However, it can be transferred to the problem of reaching a different goal state in System A (near transfer). In contrast, strategy knowledge acquired in the context of System A can likely be transferred to System B. This is particularly plausible when the strategy refers to the acquisition of structural knowledge. For example, if participants learn to apply the VOTAT strategy to System A successfully, we expect them to try it also when confronted with a new System B. Such cross-situational relevance has been shown repeatedly for the VOTAT strategy (Müller et al., 2013; Wüstenberg et al., 2014). We indicate the fact that VOTAT can be applied to a wide range of problems by referring to it as a comprehensive strategy.

Further assumptions of our preliminary process model pertain to the role of working memory (WM). We assume that the various strategies that serve knowledge acquisition are differing with respect to WM requirements. A simple trial and error strategy, associated with low WM load, is not efficient for learning the causal structure of a system, but may be suitable for acquiring I-O knowledge—which is probably often memorized implicitly (Dienes and Fahey, 1998; Hundertmark et al., 2015). The VOTAT strategy on the other hand puts a heavy load on WM and is suitable for acquiring structural knowledge. To substantiate

such assumptions, we adopt the terminology of cognitive load theory (Sweller, 1988; Sweller and Chandler, 1994). Solving a new complex problem yields intrinsic cognitive load. Corbalan et al. (2006) describe this to the point: “In terms of cognitive load theory the difficulty of a task yields intrinsic cognitive load, which is a direct result of the complex nature of the learning material. That is, intrinsic cognitive load is higher when the elements of the learning material are highly interconnected (...) and lower when they are less interconnected” (p. 404). Cognitive load associated with learning is called “germane load.” As the capacity of WM is limited, high intrinsic load leaves little capacity for germane load, thus leading to poor learning. Together, these assumptions predict that the difficulty and complexity of a source problem restrain the learning of generalizable knowledge about structures or strategies, leading to poor transfer. This prediction has been confirmed by Vollmeyer et al. (1996) in the context of CPS.

In summary, to learn comprehensive strategies such as VOTAT, learning opportunities should not be too complex. We suppose that transfer experiments are particularly useful for investigating the reach or comprehensiveness of knowledge elements and strategies.

To test some of the predictions of our preliminary process model, we have developed Dynamis2, a new CPS environment that accentuates the aspect of dynamics, which has been central in early work on CPS (e.g., Dörner and Schaub, 1994). Like MicroDYN, it is based on Funke’s (1991, 1993) Dynamis approach, which uses linear equations for calculating state changes of the system’s variables. Unlike the traditional approach, Dynamis2 simulates system dynamics in real time, which means that the state of the system is mandatorily updated every second.

The user can apply inputs at any time. A typical run with Dynamis2 comprises 250 simulation steps. As much of the research with CDC tasks has been done with systems whose states are updated in less than 9 time steps—triggered by the user—we regard Dynamis2 as an important step toward investigating dynamic decision making that deserves this label (cf. Fischer et al., 2015b; Schoppek and Fischer, 2015).

The primary goal of the present study was to test assumptions about the transfer of knowledge elements, in particular strategic knowledge, from one CDC task to another. We did this with a transfer experiment where the source function and the target function of two CPS environments were completely combined<sup>3</sup>. This enabled us to estimate transfer effects in both directions. Secondary goals were to explore the psychometric properties of Dynamis2, and to use it as validation criterion for the more established MicroDYN (Greiff et al., 2012). MicroDYN has not been validated extensively with other standardized CPS tasks (but see Greiff et al., 2013, 2015; Neubert et al., 2015). Therefore, it appears worthwhile to test the expectation that MicroDYN predicts performance in Dynamis2 over and above intelligence. In a fashion that was common at the time when we planned the experiment, we used figural reasoning as a proxy for general intelligence. We will discuss the implication of this decision and its relation to recent findings about broader operationalizations of intelligence in the discussion section (Kretschmar et al., 2016; Lotz et al., 2016).

We expected (1) positive transfer from MicroDYN to Dynamis2, mediated by the VOTAT strategy. As demonstrated by Wüstenberg et al. (2014), the extent of using this strategy predicts performance in MicroDYN. As VOTAT was in the focus of discussion about strategies in CDC tasks at the time when we designed the experiment, we did not explicitly expect PULSE as a mediator. However, we investigated the role of that strategy in *post-hoc* analyses. We expected (2) less to no transfer from Dynamis2 to MicroDYN, because the former is more difficult than the latter. Due to the quick time lapse of Dynamis2, the learner has to coordinate several concurrent subtasks in real time: Observing the course of the system, analyzing the effects of their actions, and planning new interventions. In terms of cognitive load theory (Sweller, 1988; Sweller and Chandler, 1994), this results in much more intrinsic cognitive load than MicroDYN, where the environment guides the course of action. Therefore, controlling Dynamis2 leaves less WM capacity open for germane load, which is necessary for conscious learning (Rey and Fischer, 2013). Based on recent evidence on the relation between CPS and intelligence (Wüstenberg et al., 2012; Greiff et al., 2013), we expected (3) that figural reasoning and MicroDYN should predict performance in Dynamis2. MicroDYN should explain unique variance in Dynamis2 (beyond figural reasoning) due to similar requirements (linear equation systems, knowledge acquisition, knowledge application).

<sup>3</sup>Problems that are used for learning in a transfer design are called source problems; problems in the transfer phase are called target problems. Combining refers to the fact that all levels of the factors “CPS environment” and “Function (source vs. target)” were combined.

## METHODS

We first introduce the instruments and the tasks we used in the experiment, including the measures for performance and proceeding, followed by the description of the design, the participants, and the procedure. Although some of the measures were only subject to exploratory analyses, which we conducted after testing the hypotheses, we report their operationalization here.

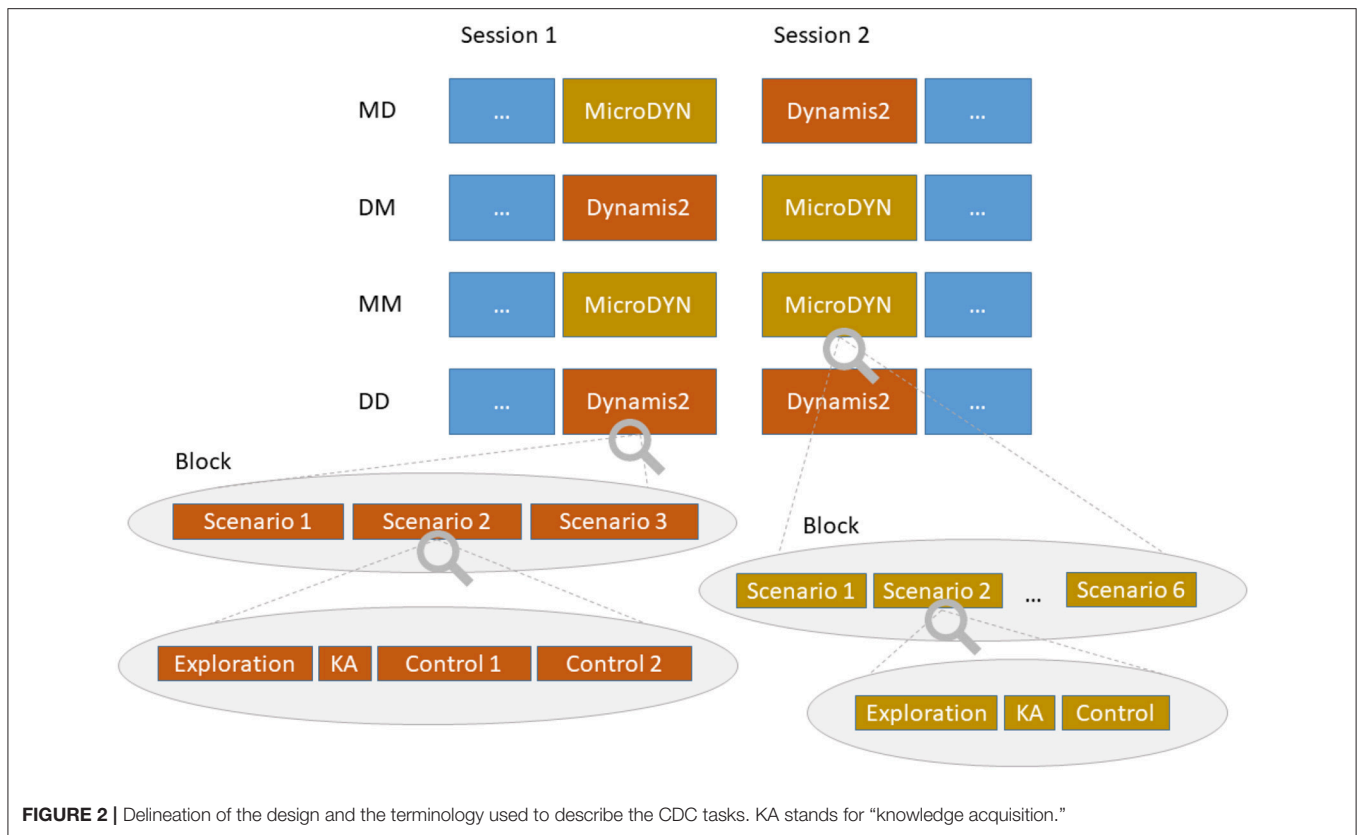
Figural reasoning was measured with a modified version of the WMT (“Wiener Matrizentest”, Formann et al., 2011). Because the original test was constructed for adolescents, we replaced two items of the original test by four more difficult items from the original APM (Raven et al., 1994). The highest possible score was 20 points. Although matrix tests load high on general intelligence assessed with broader batteries (Johnson and Bouchard, 2005), we refer to our measure as “figural reasoning”.

**Wason task:** This task requires interactive hypothesis testing (Wason, 1960). Participants are shown a list of three numbers and are asked to find out the rule that underlies the list. For example, if the list is “2 4 6,” the rule might be “three ascending even numbers” or simply “three different numbers.” To test their hypotheses, participants enter new lists and are given feedback whether the lists conform to the rule or not. To solve problems of this kind, it is important to try to falsify one’s hypotheses. Many subjects fail the task because they focus on confirming their hypotheses (Gorman and Gorman, 1984). We presented the task with three different rules. (The first was the original rule used by (Wason, 1960): “any ascending sequence”. AF devised the other two rules in the style of the first rule). As a performance measure (“Wason score”) we used the number of correctly identified rules.

## Complex Dynamic Control Tasks

Both CDC tasks we used in the experiment are based on linear equation systems with up to three input variables and up to three output variables (cf. Fischer et al., 2015a). The state of the system is calculated in discrete time steps as a function of the current state of the input variables and the state of the output variables from the preceding time step. We refer to these time steps as cycles. **Figure 2** shows an overview of the terminology we used to describe the CDC tasks. Details about the individual systems are reported in the Appendix.

**MicroDYN:** This CDC is constructed in the style of a test, consisting of several scenarios. Each scenario is defined by a specific equation system and a corresponding cover story. The process of working on the task is the same for each scenario: First, the problem solver has to explore the system’s causal structure by repeatedly varying the input variables and monitoring the effects (knowledge acquisition). To complete a cycle and see the effect of their actions the problem solver has to click a button (labeled “apply”). The problem solvers enter their insights about the systems as arrows in a causal diagram. There is a time restriction of 180 s for the exploration phase of each item. After this, the problem solvers are given goal states for each output variable that they must achieve within 90 s by manipulating the input variables up to four cycles in a row.



To assess structural knowledge, we had participants draw arrows in a causal diagram at the bottom of the screen. An arrow represented an assumed causal relation. A causal diagram was rated correct if it contained all causal relations of the system and no relation that was not simulated. Structural knowledge in the knowledge acquisition phase was scored by summing up the ternary graded degree of correctness over all causal diagrams (0: more than one error, 1: one error, 2: no errors).

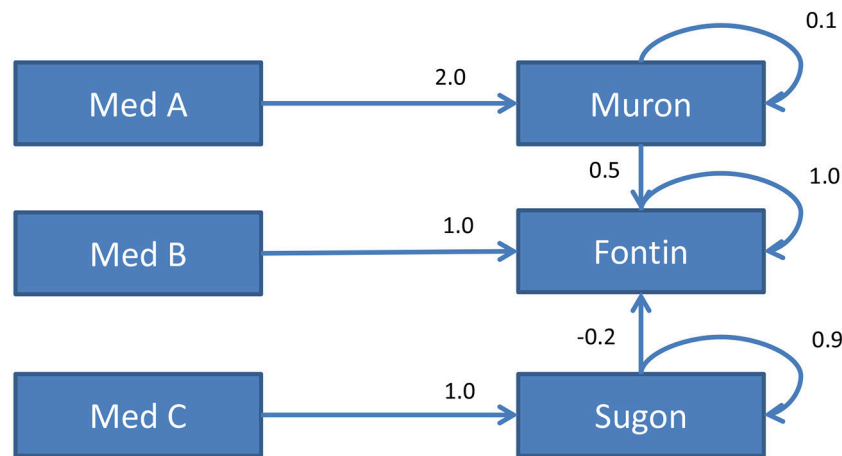
Performance in the knowledge application phase was scored by summing up the ternary graded degree of target achievement across the six items (0: targets missed, 1: targets partially met, 2: targets totally met; a single target was coded as met when the deviation was no larger than  $\pm 1$ ). As an overall performance measure, we added the knowledge acquisition score and the knowledge application score and divided the sum by two.

To determine the problem solvers' strategies, we analyzed the log files. For each cycle we observed if all input-variables were set back to zero (PULSE strategy, see below). If only one variable was set to a value different from zero at least once (VOTAT strategy), it was determined for which variable this was the case. Over all cycles of the exploration phase, we scored the proportion of input variables for which the VOTAT strategy was applied, and whether or not the PULSE strategy was applied at least once (0–1). These values were averaged across the scenarios to represent the extent of using each strategy. For example, when there are three input variables in a system and the participant

used VOTAT for two of them at least once, the VOTAT measure is 0.66.

*Dynamis2* was developed in order to emphasize the dynamic aspect of complex problem solving (Schoppek and Fischer, 2015). Like in the original *Dynamis* approach (Funke, 1991, 1993), the systems are simulated using sets of linear equations. The crucial difference is that *Dynamis2* is real-time driven, which means that the simulation is updated every second, regardless if the subject manipulates the input variables or not. This makes the dynamics of the simulated systems more tangible than in extant CPS environments such as the business microworld Tailorshop, *MicroDYN*, *Genetics Lab*, *Cherry Tree* (Beckmann and Goode, 2014), etc. In addition, genuine time pressure results for the subjects. **Figure 3** shows the causal diagram of one of the systems used in the experiment. Subjects can manipulate the three medicines Med A, Med B, and Med C (input variables) in order to control the blood values of three fictitious substances Muron, Fontin, and Sugon (output variables). Interventions can be entered for one or more input variables and applied at any time by clicking the “apply” button. Each scenario of *Dynamis2* consists of a run (250 cycles) of free exploration, followed by two runs where subjects are asked to reach and maintain a given goal state (e.g., Muron = 100, Fontin = 1,000). Performance in the goal runs is measured by goal deviation according to Equation 1, where  $n$  is the number of cycles (here 250),  $k$  is the number of goal variables,  $x_{ij}$  is the value of variable  $j$  in cycle  $i$ ,  $g_j$  is the goal value of variable  $j$ , and  $s$  is the cycle when the learners entered





**FIGURE 3** | Diagram of the causal structure of one of the Dynamis2 systems used in the experiment. The numbers denote coefficients in the linear equations that determine the state of each output variable. For example, the state of Muron at time  $t$  is given by  $Muron_t = 0.1 \cdot Muron_{t-1} + 2.0 \cdot MedA_t$ .

their first input.

$$dev = \ln \left( \frac{\sum_{i=s}^n \frac{\sum_{j=1}^k |x_{ij} - g_j|}{\sum_{j=1}^k g_j}}{\sum_{j=1}^k g_j} \right) \quad (1)$$

Because this measure is hard to interpret, we centered it on the grand mean and reversed the scale. The resulting score thus has the same orientation as the other performance measures: Higher values represent better performance.

After completion of the exploration phase, we had participants draw arrows in diagrams on paper. As a measure of structural knowledge, we subtracted the number of wrongly drawn relations from the number of correctly drawn relations and divided the difference by the number of all possible relations.

As a measure of strategy, we assessed VOTAT analogously to MicroDYN. A VOTAT event in Dynamis2 was defined by the manipulation of a single input variable, followed by at least five cycles (i.e., seconds) with no interventions. For a comprehensive measure of using the strategy, we calculated the proportion of input variables for which the VOTAT strategy was applied at least once in the exploration phases of each of the three scenarios. We averaged these proportions across the scenarios. Likewise, we defined a PULSE event by setting all input variables (back) to zero for at least five cycles and counted these events over all exploration runs. The reason why the operationalizations of PULSE differ between the two CDC tasks is that the scenarios in Dynamis2 are much longer than in MicroDYN. Due to the higher difficulty of Dynamis2 scenarios (longer runs, more dynamics), it can be quite reasonable to repeat PULSE interventions, for example to test hypotheses or to help memorizing certain effects.

## Design

We used a transfer design that allowed estimating transfer effects in both directions. As can be seen in **Figure 2**, there were four experimental conditions. In two conditions, subjects had two blocks of either MicroDYN (condition MM) or Dynamis2

(condition DD). Block 1 in these conditions consisted of separate items that were not incorporated into the calculation of transfer effects. A third condition had one block of MicroDYN, followed by one block of Dynamis2 (condition MD). The fourth condition started with one block of Dynamis2, followed by one block of MicroDYN (condition DM). Participants were randomly assigned to one of the four conditions.

In MD, DM, and the second block in MM we applied six MicroDYN scenarios. In the first block of MM, the first scenario was declared as practice scenario. All blocks of Dynamis2 consisted of three scenarios (with a different set of scenarios in the first block of DD).

## Participants

One hundred-sixty-five subjects participated in the experiment. Students of diverse majors were recruited from the University of Heidelberg ( $n = 83$ ) and from the University of Bayreuth ( $n = 82$ ). Ethical approval was not required for this study in accordance with the national and institutional guidelines. Participation was in full freedom using informed consent.

We excluded three cases from the dataset due to dubious behavior during the experiment (not complying with the instructions; aborting the experiment). In other three cases, we imputed missing values of the variables Dynamis2 score or MicroDYN score. We applied multiple regression imputation based on the cases in the respective condition. The resulting dataset comprised  $N = 162$  cases, 40 in the DD condition, 41 in DM, 42 in MD, and 39 in MM. The four conditions did not differ in figural reasoning, age and sex (all  $F_s < 1$ ).

## Procedure

The experiment took place in two sessions. Session 1 began with a short introduction and the administration of the figural reasoning test with paper and pencil. Next, subjects worked on the three items of the Wason task. Session 1 ended with the first block of complex problem solving tasks, according to the design: either six items MicroDYN or three scenarios Dynamis2 (with

two performance scores each). In Session 2, which took place 2 days after Session 1, we administered the second block of complex problem solving tasks, followed by two other tasks that are not reported in the present paper (a computerized in-basket task and an item from the wisdom questionnaire by Staudinger and Baltes, 1996). Each session lasted about 90 min.

## RESULTS

For the statistical analyses, we used an alpha level of 0.05. In addition to the significance levels, we report Cohen's (1988) effect sizes or partial  $\eta^2$ . The sample size was adequate for detecting at least medium-sized effects ( $d = 0.5$ ) with a power of 0.72 for simple mean comparisons and a power of 0.68 for one-way ANOVA (Paul et al., 2007). Descriptive statistics of the most important variables are shown in **Table 2**.

To assess the reliability of the CPS measures, we calculated Cronbach's alpha values using the results of individual scenarios as items. We obtained  $\alpha = 0.70$  for the MicroDYN score (6 items), and  $\alpha = 0.64$  for the Dynamis2 score (6 items). The measure for figural reasoning, assessed with the extended WMT, yielded  $\alpha = 0.75$ .

**Figure 4A** shows the means of the Dynamis2 scores in the three conditions that involved Dynamis2 (error bars denote 95% confidence intervals). The value in the DD group denotes performance in Block 2. We found an overall effect of condition [ $F_{(2, 121)} = 9.11, p < 0.001$ , partial  $\eta^2 = 0.132$ ], with performance linearly increasing from the DM group to the DD group. A planned comparison between the DM and the MD group yielded a significant advantage of the MD group [ $t_{(81)} = 1.82$ , one-sided  $p < 0.05$ ,  $d = 0.40$ ]. This indicates that practicing MicroDYN in Block 1 is beneficial for Dynamis2. We calculated the amount of transfer using Katona's (1940) formula (Equation 2, cited after Singley and Anderson, 1989).

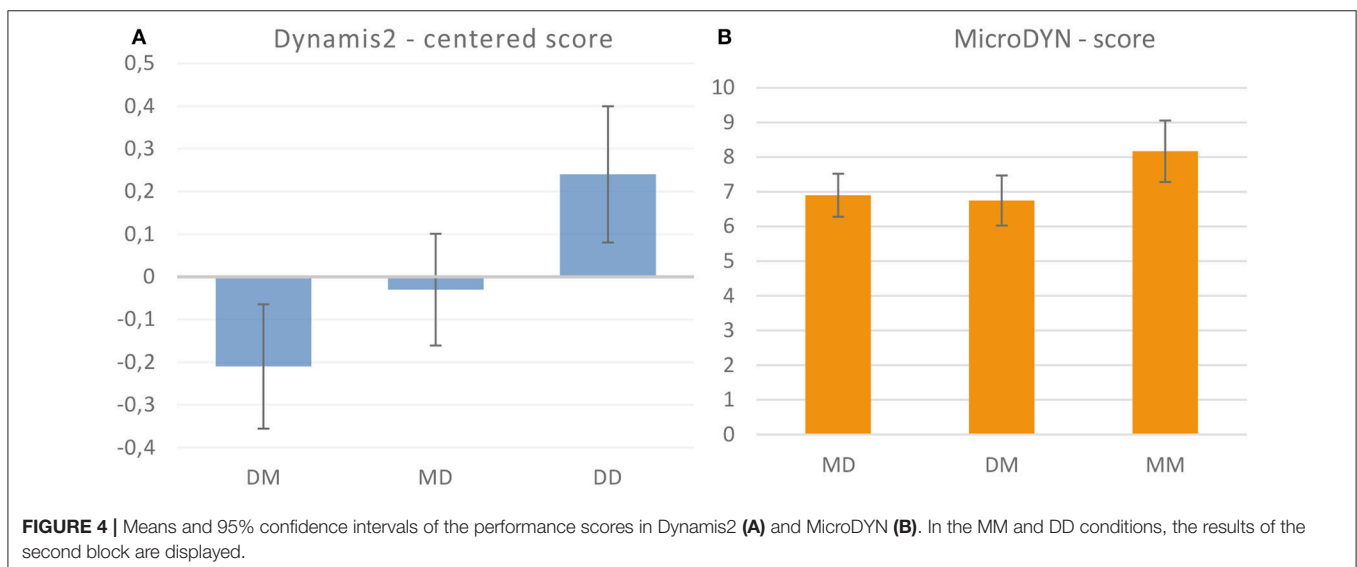
$$\%_{transfer} = \frac{C_{B1} - E_{B1}}{C_{B1} - C_{B2}} \times 100 \quad (2)$$

The denominator of Equation (2) describes the amount of improvement when the same type of problem is solved a second time ( $C$  stands for control group,  $E$  for experimental group,  $B$  for the first and second occasion). The numerator describes the difference between the baseline performance ( $C_{B1}$ ) and the performance of the experimental group in the target problem (where the experimental group has solved a different type of problem before). To estimate the transfer from MicroDYN to Dynamis, we used the mean performance in the first block of the DM group as baseline performance  $C_{B1}$ , performance in the second block of the DD group as  $C_{B2}$ , and performance in the second block of the MD group as  $E_{B1}$ . The calculation results in an estimate of 40% transfer from MicroDYN to Dynamis2. Hence, the part of Hypothesis 1 that assumed transfer is supported by the data.

A different picture emerges with the MicroDYN scores (**Figure 4B**). We found significant differences between the conditions [ $F_{(2, 120)} = 4.14, p < 0.05$ , partial  $\eta^2 = 0.065$ ], but no difference between the DM and the MD group [planned comparison,  $t_{(81)} = 0.32$ , two-sided  $p = 0.75$ ,  $d = 0.07$ ]. This means that as expected in Hypothesis 2, there is much less transfer from Dynamis2 to MicroDYN. Stating no transfer is not warranted because of the limited statistical power of our experiment.

## Prediction of Dynamis2 Performance

**Table 1** shows the bivariate correlations between the performance measures, based on pairwise deletion (i.e., the largest possible part of the sample, respectively). For example, only three fourths of the sample have worked on MicroDYN (the MD, DM, and MM groups; other three fourths have worked on Dynamis2—the MD, DM, and DD groups). We found the expected significant correlations among figural reasoning and the two CPS tasks. Performance in MicroDYN and Dynamis2 are more closely related to each other than to figural reasoning. Performance in the Wason task, which is interactive like the



CDC tasks, but not dynamic, correlates slightly, but mostly still significant with all other measures. The partial correlation between MicroDYN and Dynamis2 performance when figural reasoning is controlled for, is  $r = 0.422^{**}$ .

To analyze how MicroDYN and figural reasoning predict performance in Dynamis2 we conducted a regression analysis and a commonality analysis (see Fischer et al., 2015a). These analyses are based on the part of the sample who worked on both MicroDYN and Dynamis2 ( $n = 83$ ). Therefore, the bivariate correlation coefficients can differ from those shown in **Table 1**. The multiple regression coefficient is  $R = 0.54$ . Both predictors explain significant proportions of variance. The MicroDYN score explains a unique share of 15.4% variance ( $\beta = 0.402$ ,  $p < 0.001$ ); figural reasoning explains a unique share of 7.4% ( $\beta = 0.279$ ,  $p < 0.01$ ). The confounded variance explains 6.2% in the criterion. Altogether, these results support Hypothesis 3 that figural reasoning and MicroDYN predict performance in Dynamis2 (and that MicroDYN explains unique variance in Dynamis2, which suggests similar requirements).

## Mediation of Transfer

To test our hypothesis that transfer from MicroDYN to Dynamis2 is mediated by use of the VOTAT strategy we checked three indicators. If all three indicators are positive, the hypothesis is confirmed.

Indicator 0 is a significant correlation between the amount of using the strategy and performance in Dynamis2. This is a basic requirement that is necessary but not sufficient for demonstrating a mediation. When there is no advantage of using a certain strategy, the strategy cannot be considered to explain a transfer effect.

Indicator 1 is a significant difference of the amount of using the strategy between the MD and the DM group. When the MD

group has learned to use VOTAT in MicroDYN, then this group should use this strategy more often in Dynamis2 than the DM group who lacks this experience.

Indicator 2 provides a more challenging test of the hypothesis. It requires that there is a significant correlation between the use of the strategy in MicroDYN and performance in Dynamis2, particularly in the MD group.

As the correlation between use of VOTAT in Dynamis2 and performance in Dynamis2 is significant, but not substantial ( $r = 0.28^{**}$ ), Indicator 0 can be viewed as ambiguous and further tests will probably fail, because this indicator is essential. Indicator 1 is positive: There is a small, but significant difference in the use of the VOTAT strategy between the DM group ( $M = 0.82$ ,  $s = 0.19$ ) and the MD group [ $M = 0.89$ ,  $s = 0.16$ ,  $t_{(81)} = 1.88$ , one-sided  $p = 0.032$ ,  $d = 0.46$ ]. However, Indicator 2, the correlation between use of VOTAT in MicroDYN and performance in Dynamis2,  $r = 0.27$  (MD group), does not support the hypothesis that transfer from MicroDYN to Dynamis2 is mediated through VOTAT. Hence, the part of Hypothesis 1 that refers to attributing the transfer to the use of VOTAT is not convincingly supported by the data.

To find an explanation of the transfer effect we searched for further strategic behaviors *post-hoc*. One of them is to set one or more input variables to values greater than zero, then setting all input variables back to zero for a specified number of time steps (one in MicroDYN, five in Dynamis2). This is a useful strategy for analyzing the momentum of the output variables. We dubbed this strategy “PULSE.” For quantifying this behavior, we counted how often PULSE occurred in all exploration rounds. For that variable, all indicators to mediation were positive: The correlation between PULSE and control performance in Dynamis2 is  $r = 0.40^{**}$  (Indicator 0); there are significant differences in the use of the strategy between the relevant groups [Indicator 1: DM group:  $M = 1.85$ ,  $s = 2.47$ , MD group:  $M = 5.24$ ,  $s = 3.88$ ;  $t_{(80)} = 4.70$ ,  $p < 0.001$ ,  $d = 1.04$ ]; and also the use of PULSE in MicroDYN correlates substantially with performance in Dynamis2 (Indicator 2:  $r = 0.46^{**}$  in the MD group). So the transfer from MicroDYN to Dynamis2 can partially be explained by the fact that many subjects have learned the strategy of deploying pulses in MicroDYN and applied it successfully to Dynamis2.

**TABLE 1 |** Bivariate correlation coefficients between various performance scores ( $^*p < 0.05$ ,  $^{**}p < 0.01$ ).

	Fig. reasoning	Wason	MicroDYN
Wason	0.196* ( $n = 160$ )		
MicroDYN	0.356** ( $n = 122$ )	0.215* ( $n = 121$ )	
Dynamis2	0.342** ( $n = 123$ )	0.171 ( $n = 122$ )	0.465** ( $n = 83$ )

**TABLE 2 |** Descriptive statistics of important variables of the experiment in the four experimental conditions ( $R_t$ : theoretical range;  $R_e$ : empirical range; MDyn: MicroDYN; Dyn2: Dynamis2).

	$R_t$	MD			DM			MM			DD		
		M (s)	n	$R_e$	M (s)	n	$R_e$	M (s)	n	$R_e$	M (s)	n	$R_e$
Performance MDyn	0.0–12.0	6.90 (2.06)	42	0.0–11.0	6.75 (2.36)	41	0.5–10.5	8.17 (2.81)	39	1.5–12.0			
Performance Dyn2	–	–0.03 (0.43)	42	–0.83–1.0	–0.21 (0.48)	41	–0.96–1.1				0.24 (0.52)	40	–1.47–1.26
VOTAT MDyn	0.0–1.0	0.85 (0.21)	42	0.12–1.0	0.91 (0.16)	41	0.35–1.0	0.85 (0.24)	39	0.12–1.0			
VOTAT Dyn2	0.0–1.0	0.89 (0.16)	42	0.44–1.0	0.82 (0.19)	41	0.33–1.0				0.94 (0.11)	40	0.56–1.0
PULSE MDyn	0.0–1.0	0.70 (0.35)	42	0.0–1.0	0.61 (0.35)	41	0.0–1.0	0.70 (0.38)	39	0.0–1.0			
PULSE Dyn2	0–20+	5.24 (3.88)	42	0–14	1.85 (2.47)	40	0–8				4.02 (3.91)	40	0–15
Figural Reasoning	0–20	14.7 (3.1)	42	6–20	14.2 (2.8)	41	8–19	14.1 (4.2)	39	3–20	14.7 (3.3)	40	6–20

## Exploratory Analyses

So far, the reported results largely support our hypotheses. As we also assessed structural knowledge in Dynamis2, using structural diagrams like those in MicroDYN, we could test further predictions of the preliminary process model<sup>4</sup>. If VOTAT or PULSE are important strategies for the acquisition of structural knowledge in Dynamis2, their use should correlate with the knowledge scores in each problem.

When we aggregated the scores across the three problems, the measures are correlated in the range of  $r = 0.35^{**}$  (PULSE—knowledge) to  $r = 0.41^{**}$  (knowledge—performance). When controlling for figural reasoning, the correlations are still significant (PULSE—knowledge:  $r = 0.35^{**}$ , knowledge—performance:  $r = 0.39^{**}$ ).

When we look at the individual problems, the pattern becomes more ambiguous: The correlations between the number of PULSE events and the structural knowledge scores in three Dynamis2 problems are  $r_1 = 0.11$ ,  $r_2 = 0.34^{**}$ , and  $r_3 = 0.12$ . The correlations between structural knowledge scores and performance in these problems are  $r_1 = 0.25^{**}$ ,  $r_2 = 0.48^{**}$ , and  $r_3 = 0.16$ . So the expected role of knowledge acquisition is corroborated only in Problem 2.

This pattern of results may indicate that the low correlations in the single problems might have been due to reliability problems. However, overall this is not convincing evidence for an essential function of complete structural knowledge for performance in controlling dynamic systems. Correlations around  $r = 0.40$  involve a noticeable number of cases that do not conform to the relation suggested by the coefficient. As an example, we depict

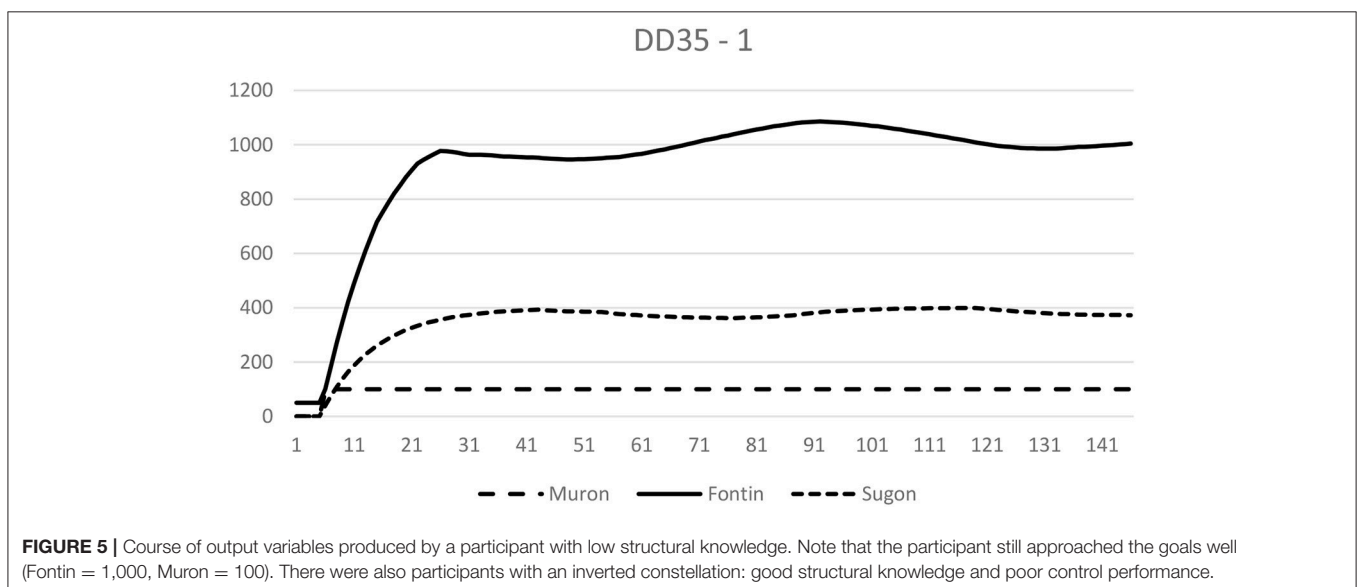
<sup>4</sup>We report this “under exploratory analyses”, because we had not put forward this hypothesis *ex ante*. In view of the current debate about false-positive results in psychological research (Pashler and Wagenmakers, 2012; Ulrich et al., 2016), we attach much importance to clearly distinguishing between the context of discovery and the context of justification.

in Figure 5 the progress of the system’s variables of a participant with low structural knowledge (standard score  $z = -1.10$ ) who nonetheless was successful in goal convergence ( $z = 1.68$ ). The goals were Fontin = 1,000 and Muron = 100.

To compare our results with studies that were published after our experiment was run (e.g., Greiff et al., 2016), we report another *post-hoc* analysis of the correlations between strategy measures and performance in both CDC tasks. VOTAT and PULSE are more closely related in MicroDYN ( $r = 0.524^{**}$ ) than in Dynamis2 ( $r = 0.330^{**}$ ). The notion that using PULSE is more significant for successful problem solving in Dynamis2 than in MicroDYN is supported by the fact that the partial correlation between PULSE and performance in Dynamis2 controlling for VOTAT is only slightly lower ( $r = 0.344^{**}$ ) than the corresponding bivariate correlation ( $r = 0.401^{**}$ ). In MicroDYN, controlling for VOTAT changes the correlation from  $r = 0.615^{**}$  to  $r = 0.410^{**}$ .

## DISCUSSION

By and large, our hypotheses are supported by the data: Performance in MicroDYN explains a unique proportion of variance in Dynamis2. We found positive transfer from MicroDYN to Dynamis2, but not in the opposite direction. This null result has to be interpreted with the reservation that the statistical power of the respective test was rather low (0.72). It may be that studies with larger samples could detect transfer effects from Dynamis2 to MicroDYN. However, the asymmetry of the transfer effects is obvious in our experiment. The assumption that transfer was mediated by using VOTAT was not clearly supported; instead, it was a different strategy called PULSE that could explain the transfer effect. PULSE is defined by setting input variables to zero and observing the system for a number of time steps ( $\geq 1$  in MicroDYN and  $\geq 5$  in Dynamis2).





This strategy—Greiff et al. (2016) refer to it as “non-interfering observation behavior”—is helpful for identifying eigendynamics (Schoppek and Fischer, 2015).

Exploratory analyses have shown that the relationships between using PULSE and the resulting structural knowledge, as well as between the latter and control performance are not as close as one might expect. Only when the respective scores were aggregated, we found substantial correlations.

With regard to aggregated results, our findings can be interpreted as supporting the standard model of CPS (Fischer et al., 2012), which assigns a critical role to knowledge acquisition (and strategies for acquiring knowledge) for the control of complex dynamic systems. As this has been shown before repeatedly (Funke, 1992; Osman, 2008; Greiff et al., 2012; Wüstenberg et al., 2012), we also want to discuss the controversial details and limitations of our findings later on. Another positive statement is that MicroDYN was successfully validated. Explaining a unique proportion of 15.4% variance in Dynamis2 performance is a considerable accomplishment, given the differences between these two classes of problems: More dynamics and momentum in Dynamis2, real-time vs. user-controlled course of events, 250 vs. on average 8 time steps (median). Also, consider the fact that the measures in the present study are manifest variables, whereas many comparable studies report proportions based on latent variables, which raises the amount of explained variance. For example, with regard to latent variables Greiff et al. (2015) report a variance overlap of 24% between MicroDYN and MicroFIN after partialling out figural reasoning (MicroFIN is another class of minimal complex systems, based on finite automata, but administered in a way similar to MicroDYN, cf. Greiff et al., 2013). Between MicroDYN and Tailorshop, they report an overlap of 7%. However, the respective study has been criticized for several methodological shortcomings, such as having administered the Tailorshop inadequately, namely in one round without a separate exploration phase (Funke et al., 2017; Kretschmar, 2017). Altogether, the variance overlap between MicroDYN and Dynamis2 (on top of the variance that both tasks share with figural reasoning) fits neatly within the range of values from comparable studies.

In recent studies, it turned out that the established finding that MicroDYN explains variance in school grades over and above figural reasoning, cannot be replicated when intelligence is operationalized broadly (Kretschmar et al., 2016; Lotz et al., 2016). This casts doubt on the distinctiveness perspective that construes CPS as an ability separate from general intelligence (Kretschmar et al., 2016). However, Kretschmar et al. (2016) still found unique covariance between MicroDYN and MicroFIN not attributable to intelligence, which can be viewed as supporting the distinctiveness view. Consistent with this, we also found considerable unique covariance between the two different CDC tasks. Irrespective of the difficult question if CPS should be construed as an ability construct in its own right, our results clearly confirm the notion that figural reasoning facilitates complex problem solving.

From a practical perspective, our results suggest that MicroDYN can be used as training device for more dynamic

task environments. However, as there are numerous instances of rather ineffective CPS training (e.g., Schoppek, 2002, 2004; Kretschmar and Süß, 2015) this prediction needs to be confirmed in further studies. We shall discuss the question what kind of real life situations are modeled by MicroDYN or Dynamis2 below.

The finding that not VOTAT could explain the transfer effect from MicroDYN to Dynamis2 but the related PULSE tactic points to the plurality of potentially relevant tactics or strategies. *Post-hoc* analyses showed that our findings correspond with recent analyses by Greiff et al. (2016), who found that controlling for VOTAT substantially reduces the relation between PULSE and knowledge acquisition in MicroDYN. However, we did not find this pattern of results in Dynamis2, where PULSE plays a discrete role. We consider two possible explanations for this difference: First, whereas all Dynamis2 scenarios involved eigendynamics, this was the case for only half of the MicroDYN scenarios (which is common practice in research with MicroDYN). Second, the real-time character of Dynamis2 makes it more obvious to vary only one variable at a time (even though it was possible to vary more variables, because the input values were transferred to the running simulation only when an apply button was pressed). Maybe a certain proportion of VOTAT events in Dynamis2 was not actually analyzed by the participants, but rather happened as a byproduct of their way of handling the CDC environment.

Findings like these raise questions about the generality of problem solving strategies: If the viability of strategies such as VOTAT and PULSE differs between certain problem classes, they could be used for classifying complex problems. Many studies have confirmed the significance of VOTAT for scientific reasoning as well as for CDC tasks from the Dynamis family (Vollmeyer et al., 1996; Chen and Klahr, 1999; Wüstenberg et al., 2014). Our results are an exception to this series, as they highlight the importance of PULSE. However, on a conceptual level the PULSE strategy is closely related to VOTAT and could be considered an extension to that strategy. On the other hand, there are many CDC tasks in- and outside the laboratory that obviously cannot be accomplished using experimental tactics like VOTAT. For example, when pilots have to handle an in-flight emergency, they are not well advised to adopt a VOTAT strategy. Generally, VOTAT is not an option in situations that forbid free exploration. In the discussion about the relationship between strategies and complex problems we should keep in mind that there are good arguments that most problem solving strategies are domain-specific to some extent (for a discussion see Tricot and Sweller, 2014; Fischer and Neubert, 2015).

Although correlations around  $r = 0.41$  (e.g., between knowledge and performance) are usually interpreted as supporting an assumed causal relation, they leave a large amount of unexplained variance, and the number of cases that differ from the general rule is not negligible. In our context, this means that there are subjects who do control our systems successfully with merely rudimentary structural knowledge. To date, most authors have taken a stand on the question about the significance of structural knowledge for performance in system control—either approving (Funke, 1992; Osman, 2008; Greiff et al., 2012;

Wüstenberg et al., 2012) or disapproving (Broadbent et al., 1986; Berry and Broadbent, 1988; Dienes and Fahey, 1998; Fum and Stocco, 2003). In our opinion, the evidence on this question is so ambiguous that an all-or-none answer is not appropriate. Some subjects seem to rely on structural knowledge, some don't (see **Figure 5**). Therefore, future research and theorizing should be aimed at specifying situational and individual conditions that predict the use (or usefulness) of structural knowledge<sup>5</sup>. As mentioned in the introduction, we believe that available working memory capacity—either varied individually or situationally (concurrent tasks, fatigue) could be such a predictor: The lower the capacity, the less promising a WM-intensive strategy is. For an excellent example of this idea applied to a static problem, see Jongman and Taatgen (1999). Although our results are consistent with these WM-related assumptions, they are not adequate for testing them directly. We plan to do this in future experiments.

If structural knowledge is not the exclusive necessary condition for successful system control, what other forms of knowledge are relevant? At this point, we can only speculate, based on our experience in the domain: Knowledge about and experience with growth and decay processes, saturation, and time delays are in our view concepts that are worth investigating. Relatedly, concepts such as wisdom may foster an appropriate way of controlling complex and dynamic systems (Fischer, 2015; Fischer and Funke, 2016).

In real life, situations where problem solvers have to find out the causal structure of a system through systematic exploration are rare. Comparable settings can be found in scientific discovery, pharmaceutical efficacy studies, organizational troubleshooting (Reed, 1997), or psychotherapy. On the other hand, there are quite a lot of situations where dynamically changing variables have to be controlled: driving a car, heating a house economically, controlling combustion processes, or monitoring vital functions in intensive care. Therefore, we consider it worthwhile to investigate how humans handle dynamic systems. However, to make our research more applicable, we—the scientific

community—should shift the focus away from questions about the acquisition of structural knowledge about simple artificial systems to questions about how humans approach more realistic CDC tasks with existing knowledge that may be limited or simplified. For example, Beckmann and Goode (2014) found that participants overly relied on their previous knowledge when dealing with a system that was embedded in a familiar context.

At last, we should not forget that although Dynamis2 exceeds MicroDYN in complexity and dynamics, both environments share some family resemblance. Therefore, we cannot generalize our results to CPS in general. Future research is necessary to investigate the common requirements of systems of the Dynamis type and more semantically rich systems such as the Tailorshop, where knowledge acquisition does not play the same role as in MicroDYN (Funke, 2014). We believe that transfer experiments could play an important role in answering these questions, too.

## ETHICS STATEMENT

This study was carried out in accordance with the recommendations of “Ethische Richtlinien der Deutschen Gesellschaft für Psychologie e.V. und des Berufsverbands Deutscher Psychologinnen und Psychologen e.V”. In accordance with the guidelines of the ethical committee at the University of Bayreuth, the study was exempt from ethical approval procedures because participation was in full freedom using informed consent and the materials and procedures were not invasive.

## AUTHOR CONTRIBUTIONS

WS planned and conducted the reported experiment together with AF. The report was written mainly by the first author, with some support from the second author, who contributed a few sections. Both authors discussed and revised the text together.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2017.02145/full#supplementary-material>

<sup>5</sup>This endeavor could well be tackled in the spirit of the early studies of Broadbent and colleagues. However, we suspect that the “salience” concept these authors focused on is closely tied to the very special characteristic of oscillatory eigendynamics, rather than a generalizable determinant of structural knowledge (see also Hundertmark et al., 2015).

## REFERENCES

- Beckmann, J. (1994). *Lernen und Komplexes Problemlösen [Learning and Complex Problem Solving]*. Bonn: Holos.
- Beckmann, J. F., and Goode, N. (2014). The benefit of being naïve and knowing it: the unfavourable impact of perceived context familiarity on learning in complex problem solving tasks. *Instruct. Sci.* 42, 271–290. doi: 10.1007/s11251-013-9280-7
- Berry, D. C., and Broadbent, D. E. (1988). Interactive tasks and the implicit-explicit distinction. *Br. J. Psychol.* 79, 251–272. doi: 10.1111/j.2044-8295.1988.tb02286.x
- Broadbent, D. E., FitzGerald, P., and Broadbent, M. H. P. (1986). Implicit and explicit knowledge in the control of complex systems. *Br. J. Psychol.* 77, 33–50. doi: 10.1111/j.2044-8295.1986.tb01979.x
- Chen, Z., and Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child Dev.* 70, 1098–1120. doi:10.1111/1467-8624.00081
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, 2.
- Corbalan, G., Kester, L., and van Merriënboer, J. J. G. (2006). Towards a personalized task selection model with shared instructional control. *Instruct. Sci.* 34, 399–422. doi: 10.1007/s11251-005-5774-2
- Dienes, Z., and Fahey, R. (1998). The role of implicit memory in controlling a dynamic system. *Q. J. Exp. Psychol. A Hum. Exp. Psychol.* 51, 593–614. doi: 10.1080/17470319808411008
- Dörner, D. (1980). On the difficulties people have in dealing with complexity. *Simul. Gaming* 11, 87–106. doi: 10.1177/104687818001100108
- Dörner, D. (1997). *The Logic of Failure: Recognizing and Avoiding Error in Complex Situations*. New York, NY: Basic Books.
- Dörner, D., and Schaub, H. (1994). Errors in planning and decision-making and the nature of human information processing. *Appl. Psychol.* 43, 433–453.
- Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G\* Power 3: a flexible statistical power analysis program for the social, behavioral, and

- biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146
- Fischer, A. (2015). Wisdom – the answer to all the question really worth asking. *Int. J. Hum. Soc. Sci.* 5, 73–83. doi: 10.11588/heidok.00019786
- Fischer, A., and Funke, J. (2016). “Entscheiden und entscheidungen: die sicht der psychologie,” in *Interdisziplinärität in Den Rechtswissenschaften. Ein Interdisziplinärer und Internationaler Dialog*, ed S. Kirste (Berlin: Duncker & Humblot), 217–229.
- Fischer, A., and Neubert, J. C. (2015). The multiple faces of complex problems: a model of problem solving competency and its implications for training and assessment. *J. Dyn. Decis. Making* 1, 1–13. doi: 10.11588/jddm.2015.1.23945
- Fischer, A., Greiff, S., and Funke, J. (2012). The process of solving complex problems. *J. Probl. Solv.* 4, 19–42. doi: 10.7771/1932-6246.1118
- Fischer, A., Greiff, S., Wüstenberg, S., Fleischer, J., Buchwald, F., and Funke, J. (2015a). Assessing analytic and interactive aspects of problem solving competency. *Learn. Individ. Differ.* 39, 172–179. doi: 10.1016/j.lindif.2015.02.008
- Fischer, A., Holt, D. V., and Funke, J. (2015b). Promoting the growing field of dynamic decision making. *J. Dyn. Decis. Making* 1, 1–3. doi: 10.11588/jddm.2015.1.23807
- Formann, A. K., Waldherr, K., and Pischwanger, K. (2011). *Wiener Matrizen-Test-2*. Göttingen: Hogrefe.
- Fum, D., and Stocco, A. (2003). “Instance vs. rule based learning in controlling a dynamic system,” in *Proceedings of the Fifth International Conference on Cognitive Modelling* (Bamberg).
- Funke, J. (1991). “Solving complex problems: human identification and control of complex systems,” in *Complex Problem Solving*, eds R. J. Sternberg and P. A. Frensch (Hove: Psychology Press), 85–222.
- Funke, J. (1992). *Wissen über dynamische Systeme. Erwerb, Repräsentation und Anwendung [Knowledge about Dynamic Systems: Acquisition, Representation, and Use]*. Heidelberg: Springer (=Lehr- und Forschungstexte Psychologie, Band 43).
- Funke, J. (1993). “Microworlds based on linear equation systems: a new approach to complex problem solving and experimental results,” in *Knowledge and Performance in Complex Problem Solving*, eds G. Strube and K. F. Wender (Amsterdam: North Holland), 313–330.
- Funke, J. (2014). Analysis of minimal complex systems and complex problem solving require different forms of causal cognition. *Front. Psychol.* 5:739. doi: 10.3389/fpsyg.2014.00739
- Funke, J., Fischer, A., and Holt, D. V. (2017). When less is less: Solving multiple simple problems is not complex problem solving—A comment on Greiff et al. (2015). *J. Intellig.* 5:5. doi: 10.3390/jintelligence5010005
- Gorman, M. E., and Gorman, M. E. (1984). A comparison of disconfirmatory, confirmatory and control strategies on Wason’s 2–4–6 task. *Q. J. Exp. Psychol.* 36, 629–648.
- Greiff, S., and Fischer, A. (2013). Der Nutzen einer komplexen Problemlösekompetenz: theoretische Überlegungen und empirische Befunde. *Z. Pädagog. Psychol.* 27, 27–39. doi: 10.1024/1010-0652/a000086
- Greiff, S., and Funke, J. (2009). “Measuring complex problem solving: the MicroDYN approach,” in *The Transition to Computer-Based Assessment - Lessons Learned from Large-Scale Surveys and Implications for Testing*, eds F. Scheuermann and J. Björnsson (Luxembourg: Office for Official Publications of the European Communities), 157–163.
- Greiff, S., and Funke, J. (2010). “Systematische Erforschung komplexer Problemlösefähigkeit anhand minimal komplexer Systeme: Projekt Dynamisches Problemlösen,” in *Zeitschrift für Pädagogik, Beiheft 56: Vol. 56. Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes*, eds E. Klieme, D. Leutner, and M. Kenk (Weinheim: Beltz), 216–227.
- Greiff, S., Fischer, A., Wüstenberg, S., Sonnleitner, P., Brunner, M., and Martin, R. (2013). A multitrait-multimethod study of assessment instruments for complex problem solving. *Intelligence* 41, 579–596. doi: 10.1016/j.intell.2013.07.012
- Greiff, S., Niepel, C., Scherer, R., and Martin, R. (2016). Understanding students’ performance in computer-based assessment of complex problem solving: an analysis of behavioral data from computer-generated log files. *Comput. Human Behav.* 61, 36–46. doi: 10.1016/j.chb.2016.02.095
- Greiff, S., Stadler, M., Sonnleitner, P., Wolff, C., and Martin, R. (2015). Sometimes less is more: comparing the validity of complex problem solving measures. *Intelligence* 50, 100–113. doi: 10.1016/j.intell.2015.02.007
- Greiff, S., Wüstenberg, S., and Funke, J. (2012). Dynamic problem solving: a new assessment perspective. *Appl. Psychol. Meas.* 36, 189–213. doi: 10.1177/0146621612439620
- Hundertmark, J., Holt, D. V., Fischer, A., Said, N., and Fischer, H. (2015). System structure and cognitive ability as predictors of performance in dynamic system control tasks. *J. Dynam. Decis. Making* 1, 1–10. doi: 10.11588/jddm.2015.1.26416
- Johnson, W., and Bouchard, T. J. (2005). The structure of human intelligence: it is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence* 33, 393–416. doi: 10.1016/j.intell.2004.12.002
- Jongman, L., and Taatgen, N. (1999). “An ACT-R model of individual differences in changes in adaptivity due to mental fatigue,” in *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society*, eds M. Hahn and S. C. Stoness (Mahwah, NJ: Erlbaum), 246–251.
- Katona, G. (1940). *Organizing and Memorizing*. New York, NY: Columbia University Press.
- Kretzschmar, A. (2017). Sometimes less is not enough: a commentary on Greiff et al. (2015). *J. Intellig.* 5, 1–9. doi: 10.3390/jintelligence5010004
- Kretzschmar, A., and Süß, H.-M. (2015). A study on the training of complex problem solving competence. *J. Dyn. Decis. Making* 1, 1–154. doi: 10.11588/jddm.2015.1.15455
- Kretzschmar, A., Neubert, J. C., Wüstenberg, S., and Greiff, S. (2016). Construct validity of complex problem solving: a comprehensive view on different facets of intelligence and school grades. *Intelligence* 54, 55–69. doi: 10.1016/j.intell.2015.11.004
- Lotz, C., Scherer, R., Greiff, S., and Sparfeldt, J. R. (2017). Intelligence in action – Effective strategic behaviors while solving complex problems. *Intelligence* 64, 98–112. doi: 10.1016/j.intell.2017.08.002
- Lotz, C., Sparfeldt, J. R., and Greiff, S. (2016). Complex problem solving in educational contexts – Still something beyond a “good g”? *Intelligence* 59, 127–138. doi: 10.1016/j.intell.2016.09.001
- Müller, J. C., Kretzschmar, A., and Greiff, S. (2013). “Exploring exploration: inquiries into exploration behavior in complex problem solving assessment,” in *Proceedings of the 6th International Conference on Educational Data Mining*, eds S. K. D’Mello, R. A. Calvo, and A. Olney (Memphis, TN: International Educational Data Mining Society).
- Neubert, J. C., Kretzschmar, A., Wüstenberg, S., and Greiff, S. (2015). Extending the assessment of complex problem solving to finite state automata. *Eur. J. Psychol. Assess.* 31, 181–194. doi: 10.1027/1015-5759/a000224
- Osman, M. (2008). Positive transfer and negative transfer/antilearning of problem-solving skills. *J. Exp. Psychol. Gen.* 137, 97–115. doi: 10.1037/0096-3445.137.1.97
- Osman, M. (2010). Controlling uncertainty: a review of human behavior in complex dynamic environments. *Psychol. Bull.* 136, 65–86. doi: 10.1037/a0017815
- Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *J. Educ. Psychol.* 84, 429–434.
- Pashler, H., and Wagenmakers, E. J. (2012). Editors’ introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspect. Psychol. Sci.* 7, 528–530. doi: 10.1177/1745691612465253
- Raven, J. C., Court, J. H., Raven, J., and Kratzmeier, H. (1994). *Advanced Progressive Matrices: [APM; RAVEN-Matrizen-Test]*. Weinheim: Beltz.
- Reed, E. N. (1997). *Organizational Troubleshooting: Asking the Right Questions, Finding the Right Answers*. Westport, CT: Quorum Books.
- Rey, G. D., and Fischer, A. (2013). The expertise reversal effect concerning instructional explanations. *Instruct. Sci.* 41, 407–429. doi: 10.1037/a0022243
- Schoppek, W. (2002). Examples, rules, and strategies in the control of dynamic systems. *Cogn. Sci. Q.* 2, 63–92.
- Schoppek, W. (2004). “Teaching structural knowledge in the control of dynamic systems: direction of causality makes a difference,” in *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, eds K. D. Forbus, D. Gentner, and T. Regier (Mahwah, NJ: Lawrence Erlbaum Associates), 1219–1224.
- Schoppek, W., and Fischer, A. (2015). Complex problem solving: single ability or complex phenomenon? *Front. Psychol.* 6:1669. doi: 10.3389/fpsyg.2015.01669

- Singley, M. K., and Anderson, J. R. (1989). *The Transfer of Cognitive Skill*. Cambridge, MA: Harvard University Press.
- Staudinger, U. M., and Baltes, P. B. (1996). Weisheit als Gegenstand psychologischer Forschung. *Psychol. Rundschau*, 47, 1–21.
- Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cogn. Sci.* 12, 257–285.
- Sweller, J., and Chandler, P. (1994). Why some material is difficult to learn. *Cogn. Instr.* 12, 185–233. doi: 10.1207/s1532690xc1203\_1
- Tricot, A., and Sweller, J. (2014). Domain-specific knowledge and why teaching generic skills does not work. *Educ. Psychol. Rev.* 26, 265–283. doi: 10.1007/s10648-013-9243-1
- Tschirgi, J. E. (1980). Sensible reasoning: a hypothesis about hypotheses. *Child Dev.* 51:1. doi: 10.2307/1129583
- Ulrich, R., Erdfelder, E., Deutsch, R., Strauß, B., Brüggemann, A., Hannover, B., et al. (2016). Inflation von falsch-positiven Befunden in der psychologischen Forschung. *Psychol. Rundschau*, 67, 163–174. doi: 10.1026/0033-3042/a000296
- Vollmeyer, R., Burns, B. D., and Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cogn. Sci.* 20, 75–100. doi: 10.1016/S0364-0213%2899%2980003-2
- Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Q. J. Exp. Psychol.* 12, 129–140.
- Wüstenberg, S., Greiff, S., and Funke, J. (2012). Complex problem solving—More than reasoning? *Intelligence* 40, 1–14. doi: 10.1016/j.intell.2011.11.003
- Wüstenberg, S., Stadler, M., Hautamäki, J., and Greiff, S. (2014). The role of strategy knowledge for the application of strategies in complex problem solving tasks. *Technol. Knowl. Learn.* 19, 127–146. doi: 10.1007/s10758-014-9222-8

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Schoppek and Fischer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Experience in a Climate Microworld: Influence of Surface and Structure Learning, Problem Difficulty, and Decision Aids in Reducing Stock-Flow Misconceptions

Medha Kumar<sup>1</sup> and Varun Dutt<sup>1,2\*</sup>

<sup>1</sup> Applied Cognitive Science Laboratory, School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, Kamand, India, <sup>2</sup> School of Humanities and Social Sciences, Indian Institute of Technology Mandi, Kamand, India

## OPEN ACCESS

### Edited by:

Annette Kluge,  
Ruhr University Bochum, Germany

### Reviewed by:

Joachim Funke,  
Universität Heidelberg, Germany  
Jon Juvina,  
Wright State University, United States  
Helen Fischer,  
Universität Heidelberg, Germany

### \*Correspondence:

Varun Dutt  
varun@iitmandi.ac.in;  
varundutt@yahoo.com

### Specialty section:

This article was submitted to  
Cognitive Science,  
a section of the journal  
Frontiers in Psychology

**Received:** 31 July 2017

**Accepted:** 22 February 2018

**Published:** 26 March 2018

### Citation:

Kumar M and Dutt V (2018)  
Experience in a Climate Microworld:  
Influence of Surface and Structure  
Learning, Problem Difficulty,  
and Decision Aids in Reducing  
Stock-Flow Misconceptions.  
Front. Psychol. 9:299.  
doi: 10.3389/fpsyg.2018.00299

Research shows that people's wait-and-see preferences for actions against climate change are a result of several factors, including cognitive misconceptions. The use of simulation tools could help reduce these misconceptions concerning Earth's climate. However, it is still unclear whether the learning in these tools is of the problem's surface features (dimensions of emissions and absorptions and cover-story used) or of the problem's structural features (how emissions and absorptions cause a change in CO<sub>2</sub> concentration under different CO<sub>2</sub> concentration scenarios). Also, little is known on how problem's difficulty in these tools (the shape of CO<sub>2</sub> concentration trajectory), as well as the use of these tools as a decision aid influences performance. The primary objective of this paper was to investigate how learning about Earth's climate via simulation tools is influenced by problem's surface and structural features, problem's difficulty, and decision aids. In experiment 1, we tested the influence of problem's surface and structural features in a simulation called Dynamic Climate Change Simulator (DCCS) on subsequent performance in a paper-and-pencil Climate Stabilization (CS) task ( $N = 100$  across four between-subject conditions). In experiment 2, we tested the effects of problem's difficulty in DCCS on subsequent performance in the CS task ( $N = 90$  across three between-subject conditions). In experiment 3, we tested the influence of DCCS as a decision aid on subsequent performance in the CS task ( $N = 60$  across two between-subject conditions). Results revealed a significant reduction in people's misconceptions in the CS task after performing in DCCS compared to when performing in CS task in the absence of DCCS. The decrease in misconceptions in the CS task was similar for both problems' surface and structural features, showing both structure and surface learning in DCCS. However, the proportion of misconceptions was similar across both simple and difficult problems, indicating the role of cognitive load to hamper learning. Finally, misconceptions were reduced when DCCS was used as a decision aid. Overall, these results highlight the role of simulation tools in alleviating climate misconceptions. We discuss the implication of using simulation tools for climate education and policymaking.

**Keywords:** stock-and-flow simulations, correlation heuristic, violation of mass balance, experience, problem structure, decision aids, heterogeneity, Dynamic Climate Change Simulator

## INTRODUCTION

Understanding stocks and flows is a fundamental process in the real world (Dörner, 1996; Sterman, 2008, 2011; Cronin et al., 2009; Fischer et al., 2015). For example, we maintain our bank accounts (a stock) as a result of our incomes (inflows) and expenses (outflows); we support our body weight (a stock) by managing our diet (inflow) and exercise (outflow); and, we maintain carbon-dioxide levels in the atmosphere (a stock) by emissions (inflow) and absorption (outflow) (Cronin et al., 2009; Dutt, 2011). Different stock-flow problems share the same underlying structure: A stock or level accumulates the inflows to it less the outflows from it (Sweeney and Sterman, 2000).

It is a well-known phenomenon that people have difficulties in understanding the dynamics of stock-flow problems (Dörner, 1996; Sterman, 2008, 2011; Cronin et al., 2009; Dutt, 2011). Stock-flow problems, even simple ones involving one stock and two flows (inflow and outflow), are difficult, even for highly educated people with strong mathematics backgrounds (Sweeney and Sterman, 2000; Sterman and Sweeney, 2002; Sterman, 2008, 2011; Cronin et al., 2009; Dutt, 2011). For example, Sweeney and Sterman (2000) presented graduate students at Massachusetts Institute of Technology with a picture of a bathtub and graphs showing the inflow and outflow of water, then asked them to sketch the trajectory of the stock of water in the tub. Although the patterns were simple, fewer than half responded correctly. We denote such difficulties in responding to stock-flow failure.

Stock-flow failure has also been documented in problems concerning Earth's climate system (Dutt, 2011). Here, people find it difficult to sketch the shape of emissions and absorptions corresponding to a carbon-dioxide (CO<sub>2</sub>) concentration trajectory. Two of the prevalent misconceptions in climate stock-flow problems are the correlation heuristic and violation of mass balance (Dutt and Gonzalez, 2012a,b). According to the correlation heuristic, people incorrectly infer that an accumulation (CO<sub>2</sub> concentration) follows the same path as the inflow (CO<sub>2</sub> emissions). This misconception assumes that stabilizing emissions would rapidly stabilize the concentration; and, emission cuts would quickly reduce the concentration and damages from climate change. This reasoning is incorrect because reliance on the correlation heuristic significantly underestimates the time delays existent between reductions in CO<sub>2</sub> emissions and their effect on the CO<sub>2</sub> concentration (Sterman, 2008; Dutt and Gonzalez, 2012a, 2013a,b; Kumar and Dutt, unpublished).

According to the second misconception in climate stock-flow problems, violation of mass balance, people incorrectly infer that atmospheric CO<sub>2</sub> concentration can be stabilized even when emissions exceed absorptions. According to mass balance violation, people think that the current state of the Earth's climate, where emissions are about double that of absorptions, would not pose a problem to future stabilization (Sterman, 2008; Dutt and Gonzalez, 2012a; Kumar and Dutt, unpublished).

Although people's wait-and-see preferences for actions against climate change are a result of several factors like social identities, party-affiliations, and denial (McCright and Dunlap, 2011),

recent research has shown that climate misconceptions are also likely to influence such preferences (Dutt, 2011). Specifically, correlation heuristic thinking leads to wait-and-see preferences because people believe that stabilizing CO<sub>2</sub> emissions is sufficient to stabilize the CO<sub>2</sub> concentration. Similarly, violation of mass balance thinking leads to wait-and-see choices because people believe that CO<sub>2</sub> concentration can be stabilized even when CO<sub>2</sub> emissions are double that of absorptions (Sterman, 2008; Dutt and Gonzalez, 2012a; Kumar and Dutt, unpublished).

Prior research has used a Climate Stabilization (CS) task to test for correlation heuristic and violation of mass balance misconceptions (Sterman and Sweeney, 2007; Sterman, 2008; Dutt and Gonzalez, 2012a,b). In the CS task, participants are given the concentration's starting value in the year 2000 and its historical trend between 1900 and 2000 on paper. Participants are asked to sketch the CO<sub>2</sub> emissions and absorptions shapes that would correspond to the projected scenario of CO<sub>2</sub> concentration between 2001 and 2100. Irrespective of educational backgrounds, people show widespread reliance on correlation heuristic and committing of violation of mass balance in their sketches in the CS task (Sterman and Sweeney, 2007; Sterman, 2008; Dutt and Gonzalez, 2012a). Overall, the CS task has been used as a measure for assessing people's stock-flow misconceptions concerning climate change (Sterman, 2008; Fischer et al., 2015).

Furthermore, recent research has documented the role that repeated feedback about cause-and-effect relationships plays on human understanding of dynamic systems, particularly for Earth's climate system (Moxnes and Saysel, 2009; Dutt and Gonzalez, 2012a). Researchers have used computer-based simulation tools and decision-making games (called microworlds) to provide repeated feedback, where a reduction in people's correlation heuristic and violation of mass balance misconceptions has been demonstrated regarding Earth's climate system (Dutt and Gonzalez, 2012a, 2013b; Kumar and Dutt, unpublished) and, dynamic systems more generally (Gonzalez et al., 2005; Gonzalez and Dutt, 2011; Dutt and Gonzalez, 2012b). For example, Dutt and Gonzalez (2012a) made participants perform in a Dynamic Climate Change Simulator (DCCS) microworld and then transferred them to the CS task immediately. Participants controlled CO<sub>2</sub> concentration to a goal level in DCCS by deciding the CO<sub>2</sub> emissions and absorptions. Next, in the CS task, participants sketched the CO<sub>2</sub> emissions and absorptions corresponding to a CO<sub>2</sub> concentration stabilization trajectory. Results revealed that exposure to DCCS before CS task reduced correlation heuristic and violation of mass balance misconceptions.

Although prior research has documented a reduction in correlation heuristic and violation of mass balance due to exposure to simulation tools, little is known on how people improve their stock-flow misconceptions when they interact with these tools. For example, Dutt and Gonzalez (2012a) gave their participants the same problem in DCCS as well as the following CS task. As the problem did not change between DCCS and CS task, it is unclear whether people learnt the structural features (how emissions and absorptions cause a change in CO<sub>2</sub> concentration under different CO<sub>2</sub> concentration scenarios) or the surface features (dimensions of emissions, absorptions, and

concentration; and, the cover-story used) of the problem in DCCS before attempting the CS task.

While performing in DCCS, one possibility is that people may learn the problem's structural features. For example, recent research has shown that structural knowledge helps people reduce their correlation heuristic and violation of mass balance misconceptions in both cases when problems encountered in the CS task are structurally similar or different compared to those presented in DCCS (Kumar and Dutt, unpublished). However, while performing in DCCS, another possibility is that people learn the surface features of the climate problem (Chi et al., 1981; Gonzalez and Wong, 2012).

In literature, procedural reinstatement principle states that performance would be better at transfer when the problems encountered during transfer are similar to those encountered during training (Healy et al., 2005). Also, heterogeneity of practice hypothesis states that training on heterogeneous (diverse) problems improves performance during transfer (Gonzalez and Madhavan, 2011). Because of the procedural reinstatement principle (Healy et al., 2005), we expect better performance when problems in the CS task (transfer) are similar in structural features or surface features to those that are learned during DCCS training (before the CS task). Also, because of heterogeneity of practice hypothesis (Gonzalez and Madhavan, 2011), we expect problems with surface or structural training during DCCS would likely produce a more efficient transfer of knowledge and improved performance in the CS task.

Moreover, as per the difficulty hypothesis, people's transfer of learning is improved when they train on difficult problems compared to easy problems (Schneider et al., 2002; Healy et al., 2005; Young et al., 2011). Thus, if people are subjected to difficult problems in DCCS, then they would likely be able to reduce their correlation heuristic and violation of mass balance misconceptions in the CS task due to the effects stated in the difficulty hypothesis. One way to create difficulty of problems in DCCS is by changing the shape of the CO<sub>2</sub> concentration curve presented: If the shape of the concentration curve is curvilinear, then this curvilinear shape would create more perceived difficulty among participants compared to when the concentration curve is straighter.

However, it is also possible that a difficult curve in DCCS may not help reduce correlation heuristic and violation of mass balance misconceptions in the CS task because of the predictions of the cognitive load theory (Sweller, 1994; De Jong, 2010). According to cognitive load theory, people possess bounded working memory capacity (Simon, 1959). Thus, if a learning task requires too much-working memory capacity, learning may get hampered (De Jong, 2010). In the DCCS task, it is possible that the processing of different elements like emission, absorption, and concentration requires certain working memory capacity. Also, the processing of the curvilinear CO<sub>2</sub> concentration curve shape may further need additional working memory capacity. Due to the overload of working memory capacity, participants may not be able to learn the stock-flow relationships in DCCS and reduce their correlation heuristic and violation of mass balance misconceptions in the CS task.

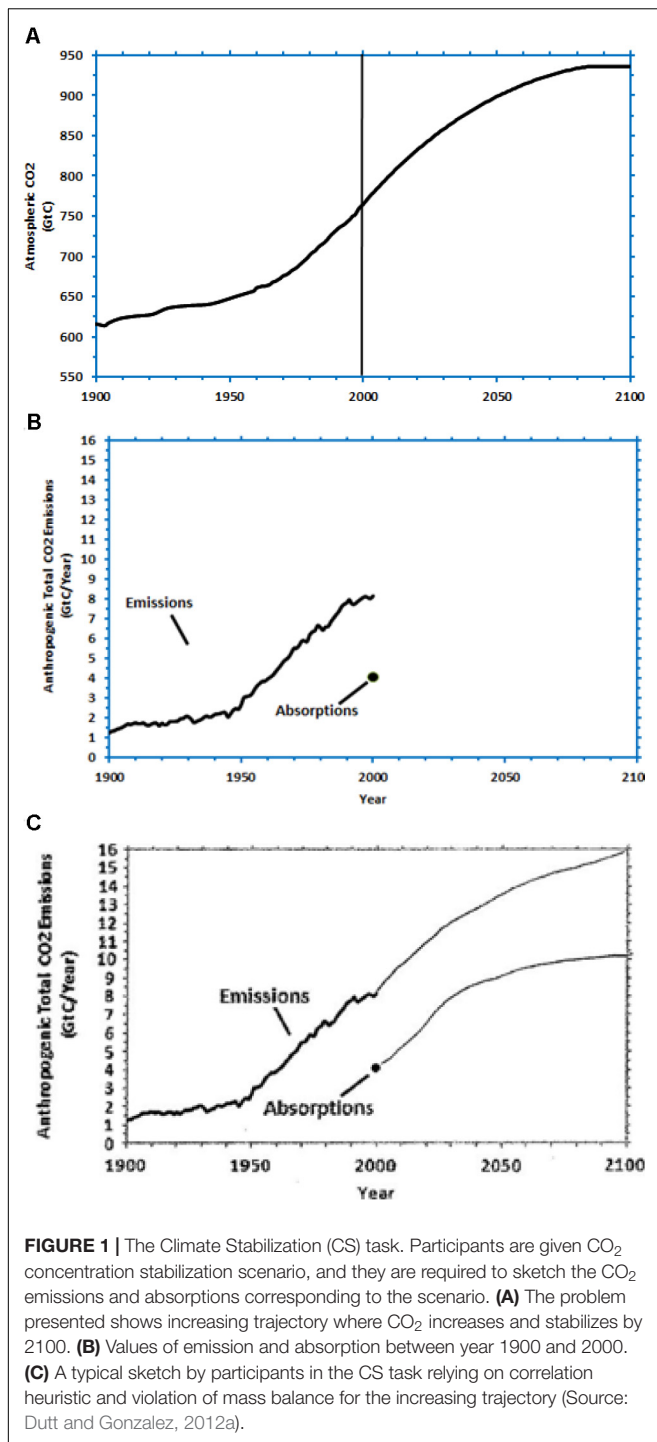
Finally, simulation tools could also be used as a side-by-side decision aid that helps people understand relationships between emissions, absorptions, and concentration by a trial-and-error procedure. There is evidence that even in simple descriptive binary-choice decision tasks, when participants are provided with experiential decision aids, they tend to rely on the experience gained in these aids in making descriptive decisions and improve their decision making (Jessup et al., 2008; Camilleri and Newell, 2011; Lejarraga and Gonzalez, 2011). When DCCS is given as aid, people are likely to get a chance to try different emissions and absorptions and see their effect on concentration. Thus, misconceptions are possible to reduce significantly when people are given an opportunity to try different values of emissions and absorptions in DCCS and to test their effect on the shape of the concentration trajectory.

The primary goal of this research is to investigate via lab-based experiments people's stock-flow misconceptions about climate change and the role that different factors like surface and structural features, problem difficulty, and decision aids play in reducing people's stock-flow misconceptions. Such research may help policymakers formulate appropriate policies for climate education in schools and colleges that make use of simulation tools to supplement conventional teaching (Meadows et al., 2016). Furthermore, this research would help provide theoretical and practical advancements in understanding the effectiveness of repeated feedback through simulation tools as an intervention in reducing misconceptions.

In what follows, we first present the background where we highlight prior research and motivate our hypotheses. Next, we report three experiments where we test how problem's surface and structural features, problem's difficulty, and decision aids help reduce misconceptions about climate change. In the first experiment, we present how problems with surface or structural training during DCCS help reduce misconceptions in the CS task. In the second experiment, we investigate how problem difficulty during DCCS training help reduce misconceptions in the CS task. In the final experiment, we study how DCCS as a decision aid helps in lowering correlation heuristic and violation of mass balance misconceptions by allowing participants to test different values of emissions and absorptions in a trial-and-error procedure. We close the paper by discussing our results and highlighting the implications of using simulation tools (like DCCS) in education and policymaking against climate change.

## BACKGROUND SECTION

Prior research in stock-flow problems concerning Earth's climate has analyzed reliance on correlation heuristic and violation of mass balance in the CS task (Serman and Sweeney, 2007; Serman, 2008; Dutt and Gonzalez, 2012a) (see **Figure 1**). In the CS task, participants are asked to sketch CO<sub>2</sub> emissions and absorptions that would stabilize the CO<sub>2</sub> concentration according to a given scenario by the year 2100 (given in **Figure 1A**). Participants are given the concentration's starting value in the year 2000 (**Figure 1B**), and its historic trends and emissions between the years 1900 and 2000. Participants are asked to sketch



the CO<sub>2</sub> emissions and absorptions shapes that would correspond to the projected scenario of CO<sub>2</sub> concentration between 2001 and 2100. **Figure 1C** shows an example of a participant that relied on correlation heuristic, whereby he inferred that the shapes of the CO<sub>2</sub> emissions and concentration should look alike. Moreover, as seen in **Figure 1C**, the participant commits violation of mass balance in her response as she fails to make emissions equal to absorption when the concentration reaches 2100. This paper uses

the CS task with different CO<sub>2</sub> concentration trajectories and cover stories to evaluate people's reliance on correlation heuristic and violation of mass balance misconceptions.

Furthermore, recent research has evaluated how repeated feedback in DCCS helps reduce correlation heuristic and violation of mass balance misconceptions (Moxnes and Saysel, 2009; Dutt and Gonzalez, 2012a,b; Kumar and Dutt, unpublished). As shown in **Figure 2**, DCCS is a dynamic replica of the CS task, it is based on a simplified and adapted climate model (Dutt and Gonzalez, 2012b), and it has been inspired by generic dynamic stocks-and-flows tasks (Gonzalez et al., 2005; Gonzalez and Dutt, 2011). In DCCS participants set yearly CO<sub>2</sub> emissions and absorptions and press "Make Decision" button. Upon pressing the "Make Decision" button, the system moves forward a certain number of years. Participants need to maintain their CO<sub>2</sub> concentration at the red goal line in the tank (which represents the atmosphere) and follow the CO<sub>2</sub> concentration trajectory shown in the bottom left panel.

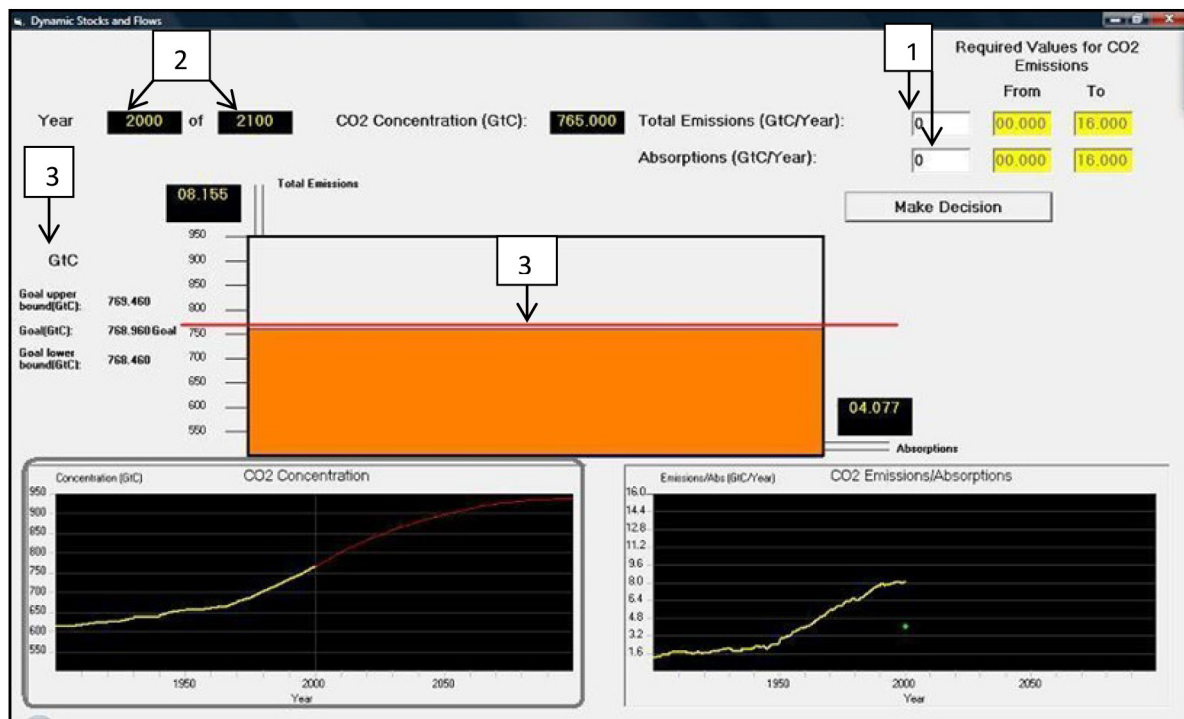
Although DCCS helps reduce people's misconceptions compared to a no-DCCS intervention (Dutt and Gonzalez, 2012a); however, little is currently known on how this reduction is influenced by problem's surface and structural features, problem's difficulty, and use of decision aids. The goal of this paper is to investigate the role of these factors in reducing people's misconceptions concerning the climate system.

First, we propose to create heterogeneous problems during DCCS training and transfer participants from DCCS training to similar/different problems in the CS task. The similarity or differences in problems between training and transfer will allow us to test participants' surface or structural learning. According to the heterogeneity of practice hypothesis (Gonzalez and Madhavan, 2011), we expect problems with surface or structural training during DCCS training to likely produce more effective transfer of knowledge and improved performance in the following CS task. Also, because of the procedural reinstatement principle (Healy et al., 2005), we expect better performance when problems in the CS task are similar in structure or surface features to those that are learned during DCCS training.

Moreover, if people are subjected to difficult problems in DCCS, then they would likely be able to reduce their correlation heuristic and violation of mass balance misconceptions in the CS task due to the difficulty hypothesis (Schneider et al., 2002; Healy et al., 2005; Young et al., 2011). However, on account of cognitive load theory and people's bounded working memory capacity (Simon, 1959; Sweller, 1994; De Jong, 2010), it is also likely that if people are subjected to difficult problems in DCCS, then they would not be able to reduce their correlation heuristic and violation of mass balance misconceptions in the CS task.

Another factor that is likely to influence people's misconceptions about climate system is the use of simulation tools as decision aids (Jessup et al., 2008; Camilleri and Newell, 2011; Lejarraga and Gonzalez, 2011). Thus, providing an experiential DCCS decision aid side-by-side to the CS task is likely to improve decision making in the CS task compared to a condition without the decision aid. In the next section, we detail experiments where we evaluated the influence of problem's surface and structural features, problem's difficulty, and use of





**FIGURE 2 |** The Dynamic Climate Change Simulator (DCCS) task. DCCS is a dynamic replica of the CS task. (1) Participants set yearly CO<sub>2</sub> emissions and absorptions and press “Make Decision” button. (2) The system now moves forward a certain number of years. (3) Participants need to maintain their CO<sub>2</sub> concentration at the red goal line in the tank (which represents the atmosphere) and follow the CO<sub>2</sub> concentration trajectory shown in the bottom left panel (Source: Dutt and Gonzalez, 2012a).

decision aids on people correlation heuristic and violation of mass balance misconceptions.

## EXPERIMENT 1: INFLUENCE OF SURFACE AND STRUCTURAL FEATURES IN REDUCING STOCK-FLOW MISCONCEPTIONS

In the first experiment, we test the influence of learning of surface and structural features in DCCS for reducing people’s misconceptions against climate change. Here, we will train people on heterogeneous problems in DCCS, which are diverse in surface and structural features. According to the heterogeneity of practice hypothesis (Gonzalez and Madhavan, 2011), one expects problems with surface or structural training during DCCS would likely produce more effective transfer of knowledge and improved performance in the CS task.

### Methods

#### Participants

Participants were recruited through an email advertisement for a climate study at Indian Institute of Technology Mandi, India. This study was carried out in accordance with the recommendations of Ethics Committee at Indian Institute of Technology Mandi with a written informed consent from all

participants. Participation was voluntary and all participants gave written informed consent before starting their study. There were 100 participants in all (74 males and 26 females). Ages ranged from 18 to 26 years (average = 21 years;  $SD = 1.5$  years). All participants were students from Science, Technology, Engineering, and Mathematics backgrounds (73% undergraduate, 19% masters, and 8% doctoral). They were randomly assigned to one of the experimental conditions involving DCCS and CS tasks. Participants were paid a flat fee of INR 50 (approximately 0.9 USD) for their participation after they completed the study.

#### Experimental Design

Participants were randomly assigned to one of four between-subjects conditions ( $N = 25^1$  in each condition): CS-Surface, CS-Structure, DCCS-Surface, and DCCS-Structure. In both DCCS-Surface and DCCS-Structure conditions, participants played 2-rounds of DCCS repeatedly with heterogeneous problems that were either based upon surface features or structural features and were then transferred to the CS task immediately. In the CS-Surface and CS-Structure conditions, participants played an unrelated task for the average time it took to complete 2-rounds in DCCS and they were then transferred to

<sup>1</sup>A power calculation with alpha level 0.05 and beta level 0.20 revealed a minimum sample size of 22. Thus, sample sizes of more than 22 were adequate for analyses reported in this paper (Faul et al., 2007)

the CS task immediately. Heterogeneity in problems was either based upon surface features or structural features.

Surface features refer to the dimensions of emissions, absorptions, and concentration; and, the cover-story used in DCCS. In the DCCS-Surface condition, participants first tackled **Figure 1**'s problem in each of the two rounds repeatedly in DCCS, however, the problem presented in each round differed randomly in the cover story and units used (i.e., in surface features). As shown in **Figure 3**, we used a glucose cover story (DCCS-Gluc; see **Figure 3A**; inflow = glucose intake, outflow = glucose metabolized, and accumulation = glucose concentration in blood over 100 time periods) and a temperature cover story (DCCS-Temp; see **Figure 3B**; inflow = heating, outflow = cooling and accumulation = temperature in a room over 100 time periods). In each of these two problems, participants controlled their accumulation trajectory in DCCS along a stabilization trajectory by making inflow and outflow decisions every 5 time periods repeatedly. After finishing two rounds in DCCS, participants were transferred to the CS task where they attempted two problems that were presented in a random order. Both these problems corresponded to **Figure 1**'s problem, where one of the problems was presented with the climate cover story (CS-Climate; i.e., just like **Figure 1**'s problem and different from problems presented during DCCS training), while the other problem was presented with the temperature cover story (CS-Temp; i.e., similar to one of the problems during the DCCS training). In both problems, participants needed to sketch the shape of inflow and outflow that corresponded to the accumulation stabilization scenario. The CS-Surface condition contained the same two problems as part of the CS task in the DCCS-Surface condition; however, the CS-Surface condition did not include DCCS training prior to the CS task. In the CS-Surface condition, participants played an unrelated Tetris game before performing in CS tasks for a duration that equaled the time taken to finish 2-rounds of DCCS performance in the DCCS-Surface condition.

Structural features refer to how emissions and absorptions cause a change in CO<sub>2</sub> concentration under different CO<sub>2</sub> concentration scenarios in DCCS. In the DCCS-Structure condition, participants first performed in two different climate problems presented randomly in DCCS. Each problem provided a different CO<sub>2</sub> stabilization trajectory, where CO<sub>2</sub> concentration increased from 765GtC in 2000 to stabilize at 936GtC by 2100 or a year before. In one of these DCCS problems, the stabilization occurred in year 2100 (**Figure 1**'s problem; DCCS-2100). In the other problem, the stabilization at 936GtC occurred much earlier in years 2070 (DCCS-2070), respectively, and the 936GtC value was maintained till the end year 2100 (see **Figure 4A** for the shape of the CO<sub>2</sub> concentration curve). In each of the two DCCS problems, participants were asked to control the CO<sub>2</sub> concentration to the stabilization trajectory over a 100-year period by making emission and absorption decisions every 5 years, repeatedly. Once participants completed 2-rounds in DCCS, they were transferred to the CS task immediately where participants attempted two problems presented in a random order. One of these two problems were **Figure 1**'s climate problem (CS-2100-Inc; i.e., like one of the problems in the

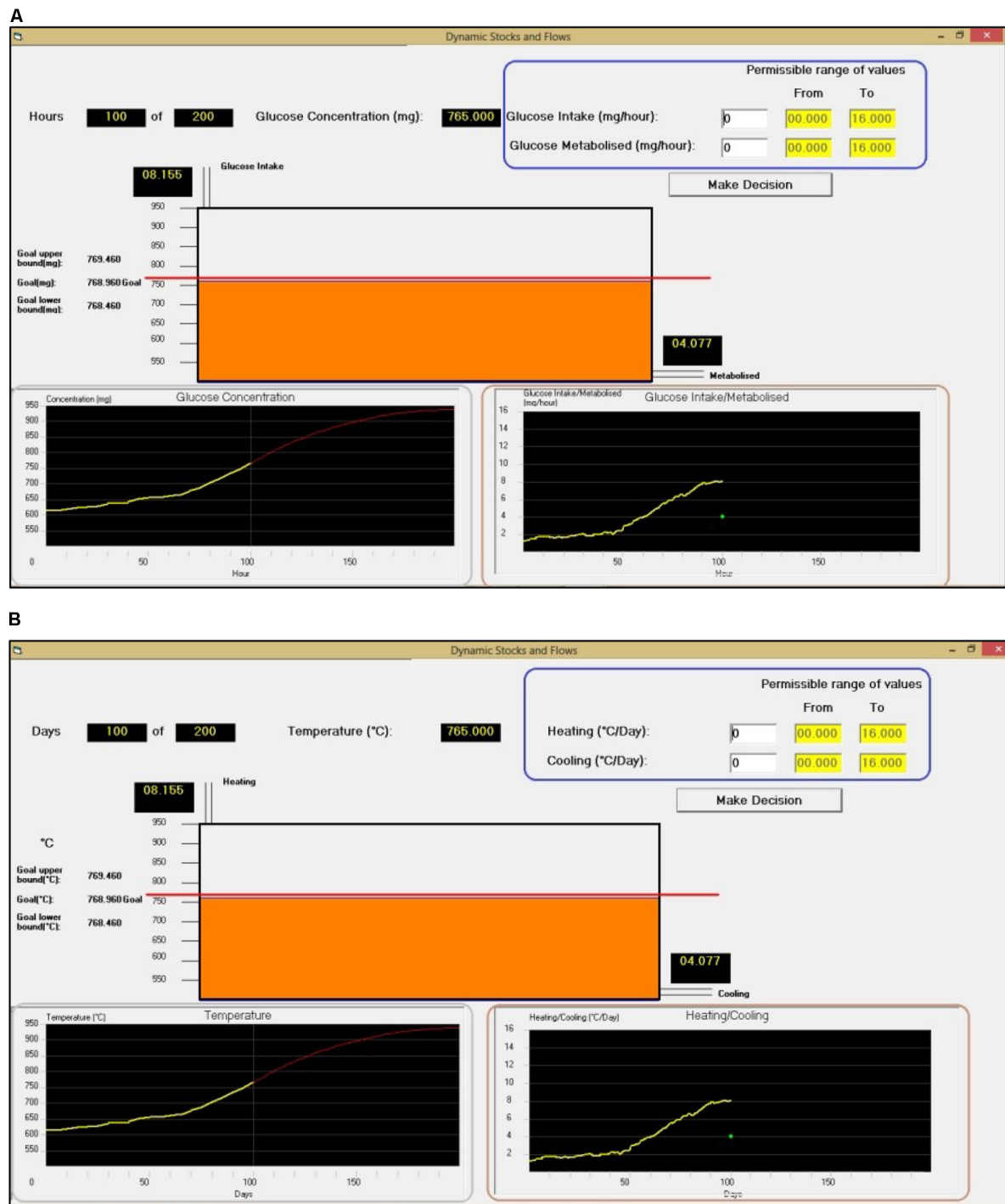
DCCS training), and the other problem was **Figure 4B**'s climate problem (CS-2100-Dec; i.e., different from all problems in the DCCS training). In both CS problems, participants needed to sketch the shape of CO<sub>2</sub> emissions and absorptions that corresponded to the CO<sub>2</sub> concentration stabilization scenario. The CS-Structure condition contained the same two problems in the CS task of the DCCS-Structure condition and did not include training in DCCS. In the CS-Structure condition, participants played an unrelated Tetris game before performing the CS task for a duration that equaled the time taken to finish 2-rounds of DCCS performance in the DCCS-Structure condition.

The CS-2100-Inc, CS-2100-Dec, CS-Temp, and CS-Climate conditions formed the control groups in the experiment. The DCCS-2070, DCCS-2100, DCCS-Temp, and DCCS-Gluc formed the training groups in the experiment. The CS-2100-Inc (DCCS), CS-2100-Dec (DCCS), CS-Temp (DCCS), and CS-Climate (DCCS) formed the test groups in the experiment.

The dependent variables were the proportion of participants relying on correlation heuristic and the proportion of participants committing violation of mass balance. A participant relied on correlation heuristic when the correlation coefficient between CO<sub>2</sub> emissions and CO<sub>2</sub> concentration during the period 2000–2100 was greater than or equal to 0.8. A participant committed violation of mass balance for the increasing trajectory stabilizing in 2100 (2070), if CO<sub>2</sub> emissions were less than CO<sub>2</sub> absorptions before year 2100 (2070) or CO<sub>2</sub> emissions were not within  $\pm 0.5\text{GtC}$  of CO<sub>2</sub> absorptions in 2100 (2070 and beyond). A participant committed violation of mass balance for the decreasing trajectory stabilizing in 2100, if CO<sub>2</sub> emissions were greater than CO<sub>2</sub> absorptions before year 2100 or CO<sub>2</sub> emissions were not within  $\pm 0.5\text{GtC}$  of CO<sub>2</sub> absorptions in 2100. Because of heterogeneity in surface or structural features in DCCS, we expected participants to possess fewer correlation heuristic and violation of mass balance misconceptions in CS conditions following DCCS compared to CS conditions without DCCS exposure. We used an alpha level of 0.05 and a power of 0.80 for our statistical analyses. The dataset for the experiment has been provided as part of **Supplementary Data Sheet S1**.

## Procedure

Participants were randomly assigned to different conditions and given instructions about the study. Participants were told about the goal that they had to achieve and they could ask clarification questions, if any, before beginning their experiment. In the DCCS-Surface and DCCS-Structure conditions, participants first performed 2-rounds in DCCS on a desktop computer and then they were transferred to CS tasks, where the CS tasks were given using a pencil-and-paper format. However, in the CS-Surface and CS-Structure conditions, participants first performed an unrelated Tetris task and then they were immediately transferred to CS tasks, which were given using a pencil-and-paper format. In the CS task, participants had to sketch CO<sub>2</sub> emissions and absorptions corresponding to the given CO<sub>2</sub> concentration trajectory. On completion of the CS task, participants were thanked and paid for their participation.



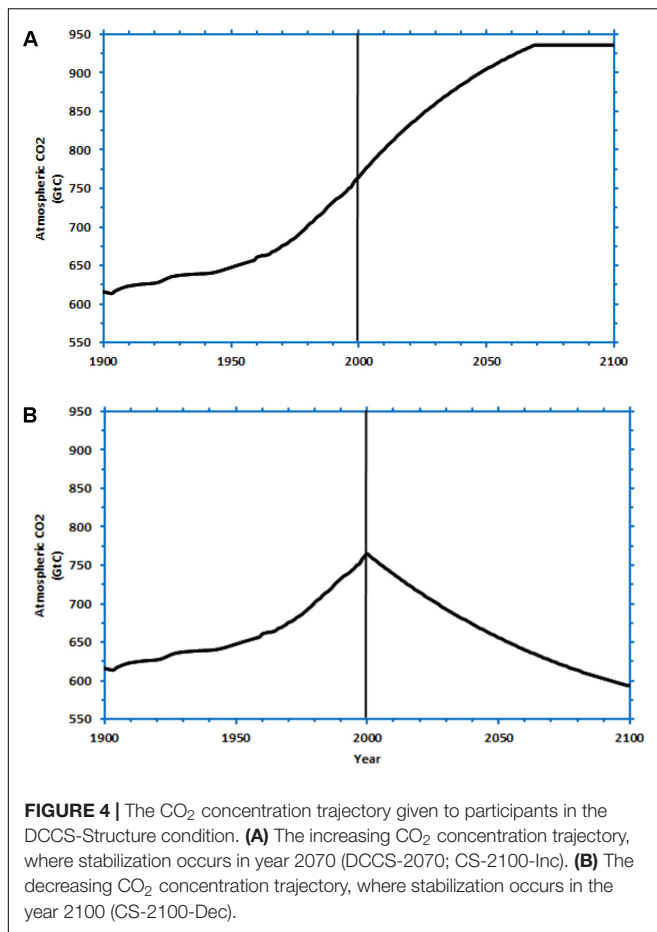
**FIGURE 3 |** Dynamic Climate Change Simulator tasks in the DCCS-Surface condition where participants need to decide the inflow and outflow values every 5 time periods such that the accumulation (Glucose Concentration or Temperature) followed the red trajectory in the bottom-left Figure. **(A)** DCCS with the glucose cover story (DCCS-Gluc). **(B)** DCCS with the temperature cover story (DCCS-Temp).

## Results

### Correlation Heuristic

We compared the correlation heuristic reliance between control groups and test groups in the structure conditions. **Figure 5**

shows the proportion of participants relying on correlation heuristic in CS tasks and DCCS in the DCCS-Structure and CS-Structure conditions. Furthermore, **Table 1** shows the comparison of different conditions and the associated inferential



statistics for correlation heuristic reliance. As seen in **Table 1**, the reliance on correlation heuristic was statistically smaller in CS-2100-Dec (DCCS) condition compared to CS-2100-Dec condition. Likewise, the reliance on correlation heuristic was statistically smaller in CS-2100-Inc (DCCS) condition compared to CS-2100-Inc condition. Furthermore, the reliance was similar in CS-2100-Dec and CS-2100-Inc conditions. Similarly, the reliance on correlation heuristic was similar in CS-2100-Dec (DCCS) task and CS-2100-Inc (DCCS) condition.

Next, we compared the correlation heuristic reliance between the control group and the test group in the surface conditions. **Figure 6** shows the proportion of participants relying on correlation heuristic in CS tasks and DCCS in the DCCS-Surface and CS-Surface conditions. As seen in **Table 1**, the reliance on correlation heuristic was statistically smaller in CS-Temp (DCCS) condition compared to CS-Temp condition. Likewise, reliance was statistically smaller in CS-Climate (DCCS) condition compared to CS-Climate condition. Furthermore, the reliance on correlation heuristic was similar in CS-Temp condition and CS-Climate condition. Similarly, reliance on correlation heuristic was similar in CS-Climate (DCCS) condition and CS-Temp (DCCS) condition.

Last, we compared the correlation heuristic reliance between control groups and test groups across the surface and structure

conditions. The reliance on correlation heuristic was statistically smaller in CS-2100-Inc condition compared to CS-Climate condition. However, the reliance on correlation heuristic was similar in CS-2100-Inc (DCCS) condition compared to CS-Climate (DCCS) condition.

Overall, in agreement with our expectations, the proportion of participants relying on correlation heuristic was statistically smaller in DCCS-Structure and DCCS-Surface conditions compared to CS-Structure and CS-Surface conditions, respectively. Also, the correlation heuristic proportions were similar in the CS-2100-Inc (DCCS) and CS-Climate (DCCS) conditions. This latter finding suggested that both the structure and surface features were similar in their ability to reduce people's correlation heuristic misconceptions.

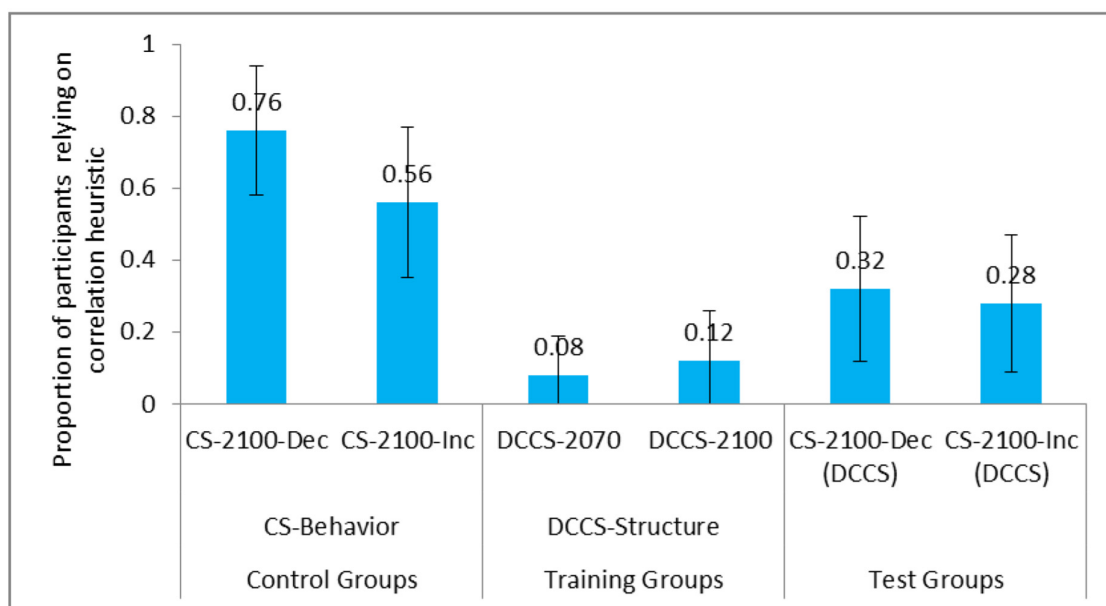
### Violation of Mass Balance

We compared the proportion of participants committing violation of mass balance between control groups and test groups across the structure conditions. **Figure 7** shows the proportion of participants committing violation of mass balance in CS tasks and DCCS in the DCCS-Structure and CS-Structure conditions. Furthermore, **Table 2** shows the comparison of different conditions and the associated inferential statistics for mass balance violation. As seen in **Table 2**, the proportion of violation of mass balance was statistically smaller in the CS-2100-Dec (DCCS) condition compared to the CS-2100-Dec condition. Furthermore, the proportion of violation of mass balance was statistically smaller in CS-2100-Inc (DCCS) condition compared to CS-2100-Inc condition. The proportion of violation of mass balance was similar in CS-2100-Inc (DCCS) condition and CS-2100-Dec (DCCS) condition. Similarly, the proportion of violation of mass balance was similar in CS-2100-Inc condition and CS-2100-Dec condition.

Next, we compared the violation of mass balance between control groups and test groups across the surface conditions. **Figure 8** shows the proportion of participants committing violation of mass balance in CS tasks and DCCS in the DCCS-Surface and CS-Surface conditions. As seen in **Table 2**, the proportion of violation of mass balance was statistically smaller in CS-Temp (DCCS) condition compared to CS-Temp condition. Likewise, the proportion of violation of mass balance was statistically smaller in CS-Climate (DCCS) condition compared to CS-Climate condition. Furthermore, the proportion of violation of mass balance was similar in the CS-Temp condition and CS-Climate condition. Similarly, the proportion of violation of mass balance was similar in CS-Climate (DCCS) condition and CS-Temp (DCCS) condition.

Last, we compared the violation of mass balance across the surface and structure conditions. The proportion of violation of mass balance was similar in CS-2100-Inc condition compared to CS-Climate condition. Similarly, the proportion of violation of mass balance was similar in CS-2100-Inc (DCCS) condition compared to CS-Climate (DCCS) condition. This latter finding suggested that both the structure and surface features were similar in their ability to reduce people's violation of mass balance misconceptions.





**FIGURE 5 |** Proportion of participants relying on correlation heuristic in CS tasks and DCCS in CS-Structure and DCCS-Structure conditions. The CS-2100-Dec (DCCS) task and CS-2100-Inc (DCCS) task refer to CS tasks following the DCCS performance in the DCCS-Structure condition. The error bars represent 95% confidence interval around the point estimate.

**TABLE 1 |** Comparison of different conditions involving correlation heuristic reliance among participants.

Condition comparisons	Statistical inference		
	$\chi^2 (1)$	$p$	$\phi$
CS-2100-Dec (DCCS) (0.32) < CS-2100-Dec (0.76)	9.74	<0.001	0.44
CS-2100-Inc (DCCS) (0.28) < CS-2100-Inc (0.56)	4.02	0.04	0.28
CS-2100-Dec (0.76) ~ CS-2100-Inc (0.56)	2.23	0.14	0.21
CS-2100-Dec (DCCS) (0.28) ~ CS-2100-Inc (DCCS) (0.32)	0.09	0.76	0.04
CS-Temp (DCCS) (0.24) < CS-Temp (0.96)	27.00	<0.001	0.73
CS-Climate (DCCS) (0.44) < CS-Climate (0.92)	13.24	<0.001	0.51
CS-Temp (0.96) ~ CS-Climate (0.92)	0.36	0.55	0.08
CS-Temp (DCCS) (0.24) ~ CS-Climate (DCCS) (0.44)	2.23	0.14	0.21
CS-2100-Inc (0.56) < CS-Climate (0.92)	8.42	<0.001	0.41
CS-2100-Inc (DCCS) (0.28) ~ CS-Climate (DCCS) (0.44)	1.39	0.24	0.17

The number in the bracket represents the proportion of participants relying on correlation heuristic. The symbol ~ indicates that the proportions in two conditions were similar to each other.

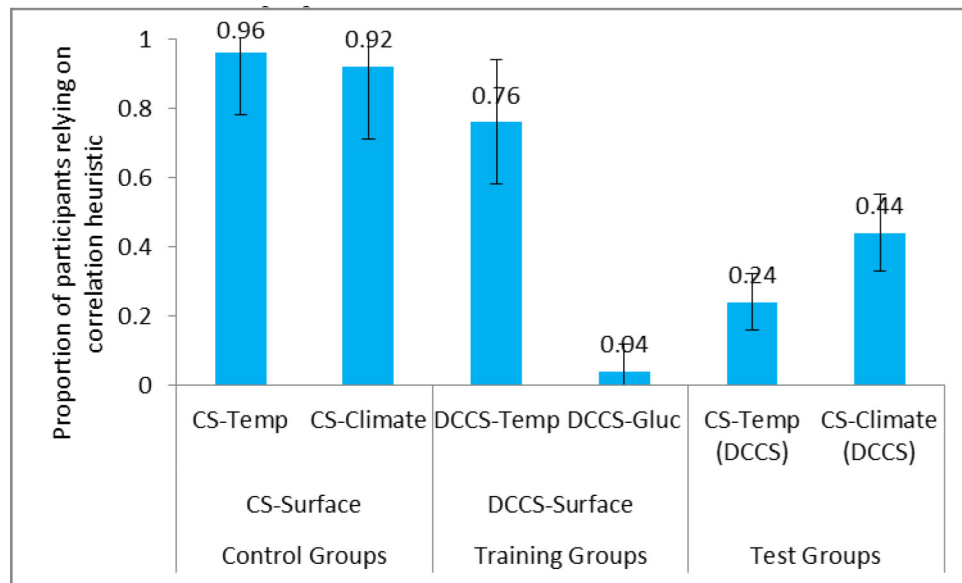
Thus, overall, the experience gained in DCCS helped participants to reduce mass balance violations. Furthermore, the violation of mass balance reduction helped participants to

perform better in the following CS task in the DCCS conditions compared to that in the CS conditions in both structure and surface condition.

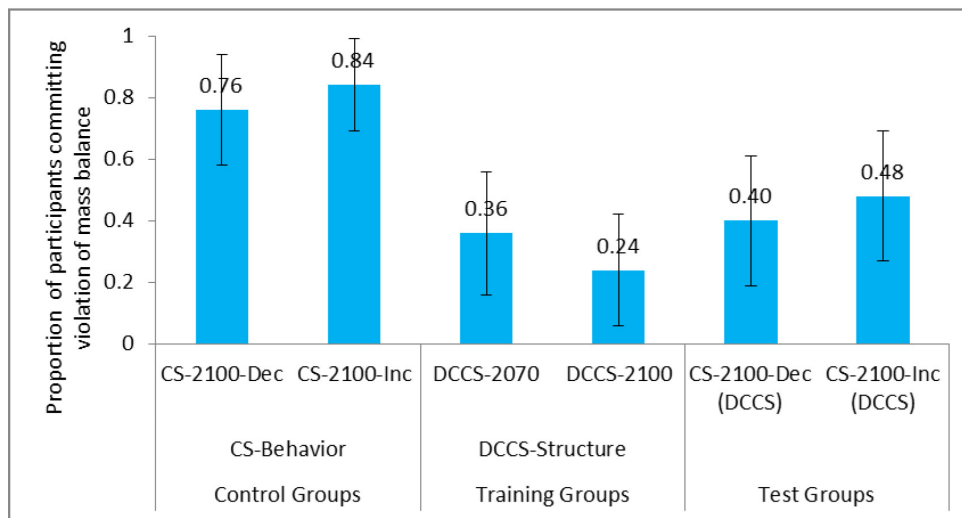
## Discussion

The comparison of the problems in the CS tasks of DCCS condition and CS condition allowed us to measure the effectiveness of the surface or structural heterogeneity in reducing correlation heuristic and violation of mass balance misconceptions. In both the surface and structure conditions, misconceptions related to correlation heuristic and violation of mass balance reduced significantly in the CS tasks following DCCS compared to CS tasks without exposure in DCCS.

First, we found that when we changed the problem's structural features between DCCS and the following CS task (i.e., change the way CO<sub>2</sub> emissions and absorptions affect the CO<sub>2</sub> concentration), misconceptions reduce significantly in the CS task post DCCS performance. This finding agrees with recent research that showed that structural knowledge helped people reduce their correlation heuristic and violation of mass balance misconceptions in both cases when problems encountered in the CS task are structurally similar or different compared to those presented in DCCS (Kumar and Dutt, unpublished). In our study, when people attempt to follow different trajectories of CO<sub>2</sub> concentration in DCCS, then this exposure to heterogeneous system dynamics likely enables them to learn that the CO<sub>2</sub> concentration increases when CO<sub>2</sub> emissions are greater than CO<sub>2</sub> absorptions, decreases when CO<sub>2</sub> emissions are smaller than CO<sub>2</sub> absorptions, and stabilizes when CO<sub>2</sub> emissions equal CO<sub>2</sub> absorptions.



**FIGURE 6 |** Proportion of participants relying on correlation heuristic in surface conditions. The CS-Temp (DCCS) and CS-Climate (DCCS) refer to CS tasks following the DCCS performance in the DCCS-Surface condition. The error bars represent 95% confidence interval around the point estimate.



**FIGURE 7 |** Proportion of participants committing violation of mass balance in CS tasks and DCCS in CS-Structure and DCCS-Structure conditions. The CS-2100-Dec (DCCS) task and CS-2100-Inc (DCCS) task refer to CS tasks following the DCCS performance in the DCCS-Structure condition. The error bars represent 95% confidence interval around the point estimate.

Second, we found that when we changed the problem's surface features in DCCS, then misconceptions also reduced significantly in the CS task post DCCS performance. One likely reason for this finding is that people get to learn via DCCS that the same system dynamics applies across different dimensions and cover stories. Thus, they could transfer this learning in CS tasks post DCCS performance.

Overall, our results agree with the procedural reinstatement principle (Healy et al., 2005), where we found improved performance when problems in the CS task were similar in

structure or surface features to those that were learned during DCCS training (prior to the CS task). Also, our results agree with the heterogeneity of practice hypothesis (Gonzalez and Madhavan, 2011), where we found that problems with surface or structural training during DCCS were able to produce more effective transfer of knowledge and improved performance in the CS task.

There were some differences in the curve shapes and cover stories used between tasks across surface and structure conditions. Thus, we could not compare all tasks across these

**TABLE 2 |** Comparison of different conditions involving violation of mass balance among participants.

Condition comparisons	Statistical inference		
	$\chi^2 (1)$	$p$	$\phi$
CS-2100-Dec (DCCS) (0.40) < CS-2100-Dec (0.76)	6.65	<0.001	0.36
CS-2100-Inc (DCCS) (0.48) < CS-2100-Inc (0.84)	7.22	<0.001	0.38
CS-2100-Dec (0.76) ~ CS-2100-Inc (0.84)	0.50	0.48	0.10
CS-2100-Dec (DCCS) (0.40) ~ CS-2100-Inc (DCCS) (0.48)	0.33	0.57	0.08
CS-Temp (DCCS) (0.20) < CS-Temp (0.92)	26.30	<0.001	0.72
CS-Climate (DCCS) (0.24) < CS-Climate (0.88)	20.78	<0.001	0.64
CS-Temp (0.92) ~ CS-Climate (0.88)	0.22	0.64	0.07
CS-Temp (DCCS) (0.20) ~ CS-Climate (DCCS) (0.24)	0.12	0.74	0.05
CS-2100-Inc (0.84) ~ CS-Climate (0.88)	0.17	0.68	0.06
CS-2100-Inc (DCCS) (0.48) ~ CS-Climate (DCCS) (0.24)	3.13	0.07	0.25

The number in the bracket represents the proportion of participants committing violation of mass balance. The symbol ~ indicates that the proportions in two conditions were similar to each other.

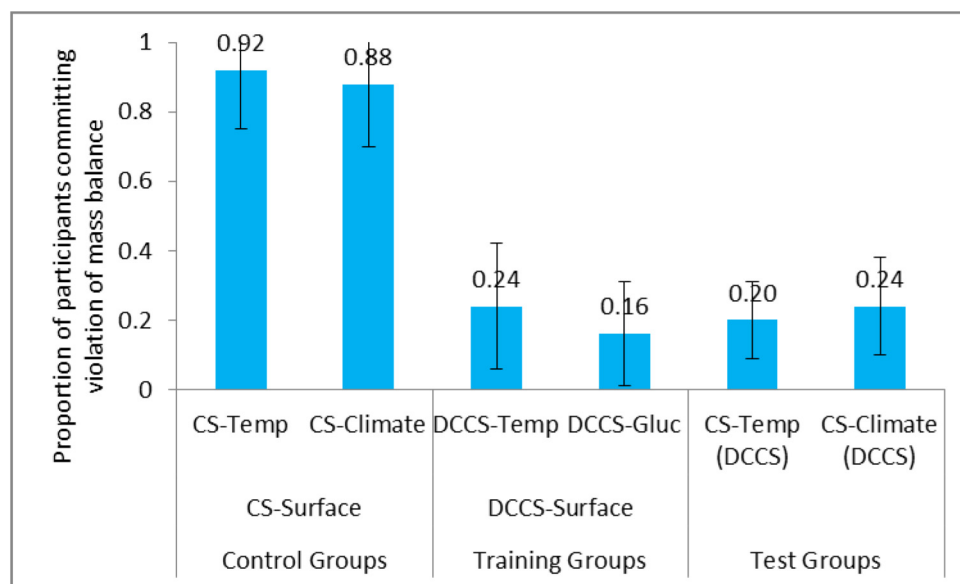
conditions. However, upon comparing tasks that were similar in their curve shapes and cover stories used, we did find a similar reduction in correlation heuristic and violation of mass balance misconceptions across the surface and structure conditions.

Overall, these results indicate that both surface and structural heterogeneity is equally powerful in reducing people's stock-flow misconceptions.

Although the problems used in the current experiment created learning of structural and surface features for participants, there may be other ways of creating effective training conditions. However, as part of future work we would like to compare structure and surface heterogeneity with homogenous conditions. For example, one other way learning could be influenced during DCCS training is by varying the difficulty level of problems in DCCS. The problem difficulty could be varied in DCCS based upon the shape of CO<sub>2</sub> concentration trajectory that participants are asked to follow in DCCS. The next experiment explores the effects of problem difficulty in reducing correlation heuristic and violation of mass balance misconceptions.

## EXPERIMENT 2: EFFECT OF DIFFICULTY OF PROBLEMS IN REDUCING STOCK-FLOW MISCONCEPTIONS

Another way in which training conditions might differ is by the difficulty of problems encountered. For example, school children may be trained on simple and difficult problems in the classroom to prepare them for different problems in their exam. According to the difficulty hypothesis (Schneider et al., 2002; Young et al., 2011), transfer performance in the CS task should improve when training is conducted using difficult climate problems in DCCS compared to simple problems. However, it is also possible that due to the predictions from cognitive load theory (Sweller, 1994; De Jong, 2010), difficult training problems in DCCS may not lead



**FIGURE 8 |** Proportion of participants committing violation of mass balance in surface conditions. The CS-Temp (DCCS) and CS-Climate (DCCS) refer to CS tasks following the DCCS performance in the DCCS-Surface condition. The error bars represent 95% confidence interval around the point estimate.

to reductions in stock-flow misconceptions compared to simple training problems.

## Methods

### Participants

Participants were recruited through an email advertisement for a climate-study at Indian Institute of Technology, Mandi, India. This study was carried out in accordance with the recommendations of Ethics Committee at Indian Institute of Technology Mandi with a written informed consent from all participants. Participation was voluntary and all participants gave written informed consent before starting their study. There were 90 participants in all (78 males and 12 females). Ages ranged from 18 to 25 years (average = 23 years;  $SD = 1.4$  years). All participants were from Science, Technology, Engineering, and Mathematics backgrounds (88% undergraduate, 9% masters, and 3% doctoral). They were randomly assigned to one of the experimental conditions involving DCCS and CS tasks. Participants were paid a flat fee of INR 50 (approximately 0.9 USD) for their participation after they completed the study.

### Experimental Design

Participants were randomly assigned to one of the following three between-subjects conditions ( $N = 30$  in each condition): DCCS-Difficult, DCCS-Easy and CS. In the DCCS-Difficult and DCCS-Easy conditions, participants first performed 1-round in DCCS and were immediately transferred to the CS task. In the DCCS-Easy and DCCS-Difficult conditions, in DCCS, participants controlled the CO<sub>2</sub> concentration to the stabilization trajectory in each round by making inflow and outflow decisions every 5 time periods repeatedly. In the DCCS-Easy condition, the DCCS used **Figure 1**'s problem. However, in the DCCS-Difficult condition, the DCCS used **Figure 9**'s problem. The shape of CO<sub>2</sub> concentration scenario in **Figure 9**'s problem was more complex compared to that in **Figure 1**'s problem (although the CO<sub>2</sub> concentration in both problems had about the same values and direction of movement over time). The complexity of the concentration curve made **Figure 9**'s problem more difficult compared to **Figure 1**'s problem. After participants finished

performing in DCCS, they were transferred to a different problem in the CS task. In the CS condition, however, participants played an unrelated Tetris task for the average time it took to complete 1-round in the DCCS task (in the conditions involving DCCS) and were transferred to the CS task immediately. In CS tasks across all conditions, participants attempted the problem shown in **Figure 4B**, where they sketched the shape of CO<sub>2</sub> emissions and absorptions that corresponded to a decreasing CO<sub>2</sub> concentration stabilization trajectory between 2001 and 2100. In this experiment, the CS condition formed the control group, the DCCS-Easy and DCCS-Difficult conditions formed the training groups, and the CS (DCCS-Easy) and CS (DCCS-Difficult) formed the test groups.

The dependent variables were the proportion of participants relying on correlation heuristic and the proportion of participants committing violation of mass balance. The coding used to classify participants as relying on correlation heuristic and committing violation of mass balance across the control, training, and test groups was the same as that used in Experiment 1. The alpha and power levels were same as reported in experiment 1. The dataset for the experiment has been provided as part of **Supplementary Data Sheet S1**.

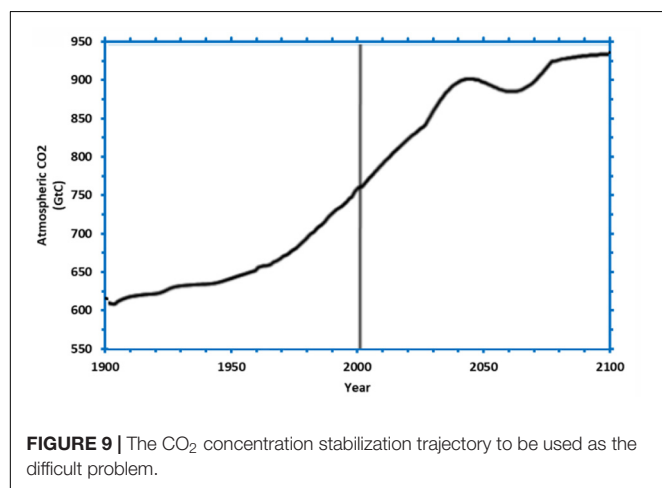
### Procedure

Participants were randomly assigned to different conditions and given instructions about the study. Participants were told about the goal that they had to achieve and they could ask clarification questions, if any, before beginning their experiment. In the DCCS-Easy and DCCS-Difficult conditions, participants performed 1-round in DCCS on a desktop computer and then they were transferred to CS tasks, where the CS tasks were given using a pencil-and-paper format. However, in the CS condition, participants first performed an unrelated Tetris task and then they were immediately transferred to the CS task, which was given using a pencil-and-paper format. In the CS task, participants had to sketch CO<sub>2</sub> emissions and absorptions corresponding to the CO<sub>2</sub> concentration trajectory. On completion of the CS task, participants were thanked and paid for their participation.

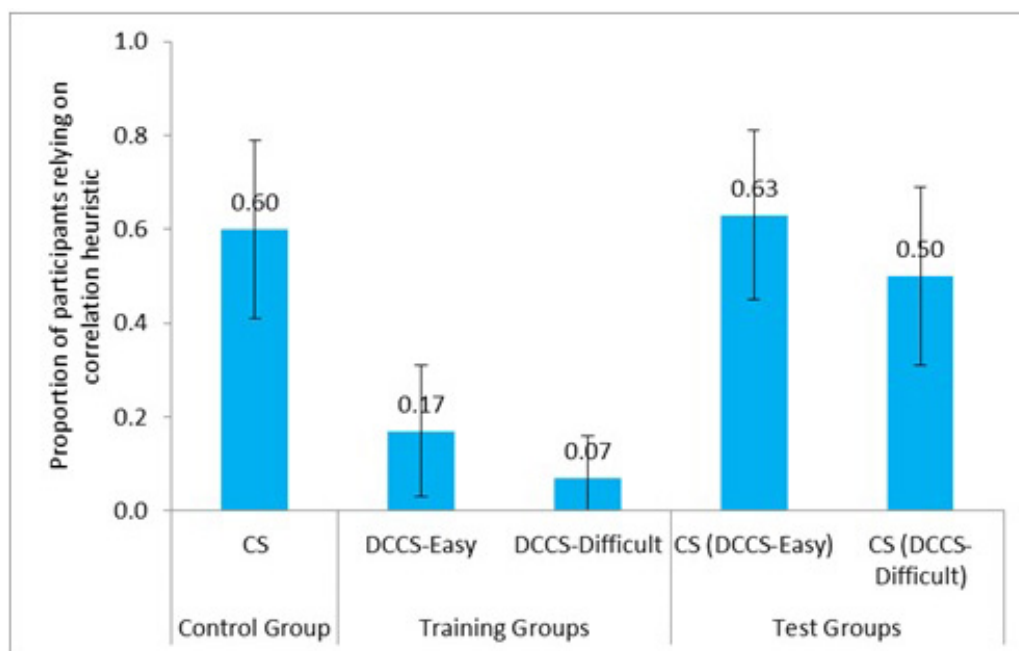
## Results

### Correlation Heuristic

We compared the correlation heuristic reliance between the control group and the test groups across the easy and difficult conditions. **Figure 10** shows the proportion of participants relying on correlation heuristic in CS tasks and DCCS in the DCCS-Easy, DCCS-Difficult, and CS conditions. Furthermore, **Table 3** shows the comparison of different conditions and the associated inferential statistics for correlation heuristic reliance. As seen in **Table 3**, the reliance on correlation heuristic was similar across the CS tasks in the CS condition and the DCCS-Difficult condition. Similarly, the reliance on correlation heuristic was similar across the CS tasks in the CS condition and the DCCS-Easy condition. Furthermore, the proportion of participants relying on correlation heuristic was similar across the CS tasks in the DCCS-Easy condition and the DCCS-Difficult condition.







**FIGURE 10 |** Proportion of participants relying on correlation heuristic in three conditions: DCCS-Easy, DCCS-Difficult, and CS conditions. The CS (DCCS-Easy) task and CS (DCCS-Difficult) task refer to CS tasks following the DCCS performance in the DCCS-Easy and DCCS-Difficult conditions. The error bars represent 95% confidence interval around the point estimate.

**TABLE 3 |** Comparison of different conditions involving correlation heuristic reliance among participants.

Condition comparisons	Statistical inference		
	$\chi^2 (1)$	$p$	$\phi$
CS (0.60) ~ CS (DCCS-Difficult) (0.50)	0.61	0.44	0.10
CS (0.60) ~ CS (DCCS-Easy) (0.63)	0.07	0.79	0.03
CS (DCCS-Easy) (0.63) ~ CS (DCCS-Difficult) (0.50)	1.09	0.29	0.13

The number in the bracket represents the proportion of participants relying on correlation heuristic. The symbol ~ indicates that the proportions in two conditions were similar to each other.

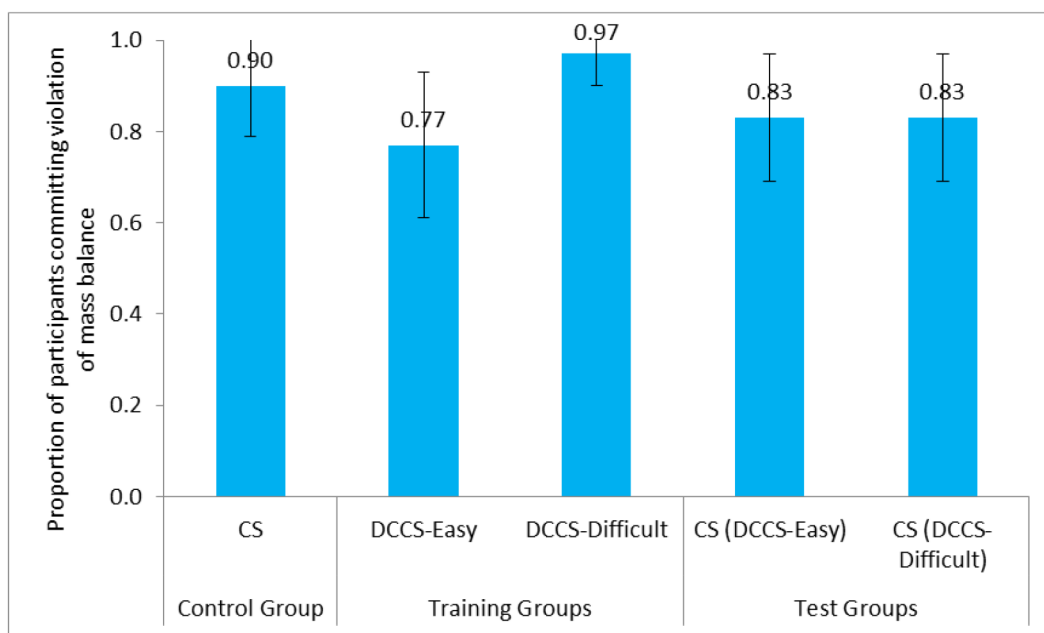
### Violation of Mass Balance

We compared the committing of violation of mass balance between the control group and the test groups across the easy and difficult conditions. **Figure 11** shows the proportion of participants committing violation of mass balance in CS tasks and DCCS in DCCS-Easy, DCCS-Difficult, and CS conditions. **Table 4** shows the comparison of different conditions and the associated inferential statistics for mass balance violation. As seen in **Table 4**, results indicated that the proportion of violation of mass balance was similar across the CS tasks in the CS condition and the DCCS-Difficult condition. Furthermore, the proportion of violation of mass balance was similar across the CS tasks in the CS condition and the DCCS-Easy condition. Likewise, the proportion of violation of mass balance was similar across the CS tasks of the DCCS-Easy condition and the DCCS-Difficult conditions.

Overall, in agreement with the expectations from cognitive load theory, the proportion of participants relying on CH and committing violation of mass balance were similar in the CS tasks of the DCCS-Difficult condition and the DCCS-Easy condition.

### Discussion

Variation in problem difficulty could be another way of enabling learning among people that reduces their stock-flow misconceptions. In this experiment, we varied problem difficulty in terms of the shape of the CO<sub>2</sub> concentration trajectory: smooth (simple) or curvilinear (difficult). We found that people could not reduce their correlation heuristic misconceptions after exposure to difficult climate problems in DCCS compared to those who were either not provided DCCS training or were only exposed to easy climate problems in DCCS. Similarly, the same intervention did not reduce the violation of mass balance misconceptions: The committing of violation of mass balance remained the same after DCCS training (among both easy and difficult problems) in the CS task compared to conditions where the CS task was given without exposure in DCCS. The lack of reduction in correlation heuristic and violation of mass balance misconceptions could be attributed to cognitive load theory (Simon, 1991; Sweller, 1994; De Jong, 2010). As per cognitive load theory, it is possible that the processing of different elements like emission, absorption, and the curvilinear concentration in the DCCS task required too much working memory capacity. Due to the cognitive overload and bounded memory capacity, participants were not able to reduce their correlation heuristic and violation of mass balance misconceptions.



**FIGURE 11 |** Proportion of participants committing violation of mass balance in three conditions: DCCS-Easy, DCCS-Difficult, and CS conditions. The CS (DCCS-Easy) task and CS (DCCS-Difficult) task refer to CS tasks following the DCCS performance in the DCCS-Easy and DCCS-Difficult conditions. The error bars represent 95% confidence interval around the point estimate.

**TABLE 4 |** Comparison of different conditions involving violation of mass balance among participants.

Condition comparisons	Statistical inference		
	$\chi^2 (1)$	$p$	$\phi$
CS (0.90) ~ CS (DCCS-Difficult) (0.83)	0.58	0.45	0.10
CS (0.90) ~ CS (DCCS-Easy) (0.83)	0.58	0.45	0.10
CS (DCCS-Easy) (0.83) ~ CS (DCCS-Difficult) (0.83)	0.00	1.00	0.00

The number in the bracket represents the proportion of participants committing violation of mass balance. The symbol ~ indicates that the proportions in two conditions were similar to each other.

Our results in this experiment did not agree with the expectations from the difficulty hypothesis (Schneider et al., 2002; Young et al., 2011). Perhaps, the shape of the difficult CO<sub>2</sub> concentration trajectory was not difficult enough in making people learn reduce their stock-flow misconceptions. Although we can only speculate currently, a more challenging CO<sub>2</sub> concentration trajectory in DCCS that gives exposure to people about increase, decrease, and stabilization of accumulation may help reduce people's misconceptions.

Beyond testing the difficulty of problems and their effectiveness in DCCS, another way for reducing correlation heuristic and violation of mass balance misconceptions could be by using simulation tools as side-by-side decision aids (e.g., a computer or calculator). The focus of the next experiment is to evaluate how DCCS could be used as a side-by-side decision aid in reducing stock-flow misconceptions.

## EXPERIMENT 3: EFFECT OF DECISION AIDS IN REDUCING STOCK-FLOW MISCONCEPTIONS

There are numerous situations in life like during schooling when students make use of decision aids (e.g., computers and calculators) to assist them in solving complex mathematical problems. Similarly, climate-scientists and climate-policymakers are likely to use decision aids (e.g., simulation tools) while formulating future greenhouse gas emission policies. For example, to evaluate the effects of future emission policies on the CO<sub>2</sub> concentrations and global temperatures we may need to rely upon decision aids. In simple descriptive binary-choice decision tasks, when participants are provided with experiential decision aids, they tend to rely on the experience gained in these aids in making descriptive decisions and improving their decision making (Jessup et al., 2008; Camilleri and Newell, 2011; Lejarraga and Gonzalez, 2011). The aim of this experiment is to evaluate the effectiveness of decision aids in reducing people's misconceptions when they have at their disposal an aid that simulates future CO<sub>2</sub> concentrations by assuming different CO<sub>2</sub> emission policies.

## Methods

### Participants

Participants were recruited through an email advertisement for a climate-study at Indian Institute of Technology Mandi, India. This study was carried out in accordance with the recommendations of Ethics Committee at Indian Institute of

Technology Mandi with a written informed consent from all participants. Participation was voluntary and all participants gave written informed consent before starting their study. There were 60 participants in all (52 males and 08 females). Ages ranged from 18 to 26 years (average = 22 years;  $SD = 1.5$  years). All participants were from Science, Technology, Engineering, and Mathematics backgrounds (85% undergraduate, 12% masters, and 3% doctoral). They were randomly assigned to one of the conditions involving DCCS and CS tasks. Participants were paid a flat fee of INR 50 (approximately 0.9 USD) for their participation after they completed the study.

## Experimental Design

Participants were randomly assigned to one of two between-subjects conditions ( $N = 30$  in each condition): Aid and No-aid. In the Aid condition, participants could use DCCS side-by-side as a decision aid while sketching the CO<sub>2</sub> emissions and absorptions in the CS task; however, in the No-aid condition, participants only sketched the CO<sub>2</sub> emissions and absorptions in the CS task and they did not use DCCS. In the Aid condition, participants could use DCCS anytime to enter 10-yearly emission and absorption values over a period of 100 years (i.e., a total of 10 values for each of the emissions and absorptions) and simulate the resulting CO<sub>2</sub> concentration. The DCCS simulated the entered emissions and absorptions rapidly within 1 to 2 seconds. Participants could then reset DCCS to the year 2000 and simulate a different set of emission and absorption values. In the Aid condition, participants could use DCCS as many times as they wanted to before they sketched the CO<sub>2</sub> emissions and absorptions in the CS task. Also, the number of times participants used the DCCS as a decision aid was recorded in the Aid condition. In the No-aid condition, participants were asked to play a Tetris game for an amount time that equaled the time that participants took to use DCCS in the Aid condition. The No-aid condition formed the control group and the Aid condition formed the test group.

The dependent variables were the proportion of participants relying on correlation heuristic and the proportion of participants committing violation of mass balance. In the Aid condition, the correlation heuristic and violation of mass balance misconceptions were analyzed in DCCS by using the averaged emission and absorption trajectory, where the average was computed across the number of times DCCS was used as a decision aid. In both Aid and No-aid conditions, participants attempted a single problem in the CS task and that was the one shown in **Figure 4B**. The coding used to classify participants as relying on correlation heuristic and committing violation of mass balance across the control and test groups was the same as that used in Experiment 1. Because of the presence of DCCS, we expected smaller proportions of correlation heuristic and violation of mass balance in the CS task in Aid condition compared to the No-aid condition. The alpha and power levels were the same as reported in experiment 1. The dataset for the experiment has been provided as part of **Supplementary Data Sheet S1**.

## Procedure

Participants were randomly assigned to different conditions and given instructions about the study. Participants were told about the goal in the CS task: to sketch the CO<sub>2</sub> emission and absorption trajectories that would correspond to the CO<sub>2</sub> concentration trajectory. Participants could ask clarification questions, if any, before starting their study. In the Aid condition, participants were encouraged to use DCCS as a decision aid side-by-side the CS task. However, in the No-aid condition, participants first performed in the unrelated Tetris task and then they were immediately transferred to the CS task. On completion of the CS task, participants were paid for their participation.

## Results

First, we analyzed the number of times DCCS was used as a decision aid in the Aid condition. Results revealed that participants used DCCS between 1 time and 7 times in the Aid condition (average = 3 times,  $SD = 1.4$  times).

### Correlation Heuristic

We compared the correlation heuristic reliance between the CS tasks across the Aid and No-aid conditions. **Figure 12** shows the proportion of participants relying on correlation heuristic in the Aid and No-aid conditions. Results revealed that reliance on correlation heuristic was statistically smaller in the CS task of Aid condition compared to the CS task of No-aid condition [ $0.30 < 0.60$ ,  $\chi^2(1) = 5.46$ ,  $p = 0.02$ ,  $\phi = 0.30$ ]. The proportion of participants relying on correlation heuristic in DCCS was close to 0.30. The correlation between the number of times DCCS was used and reliance on correlation heuristic in the CS task was small and insignificant ( $r = 0.14$ ,  $p = 0.41$ ).

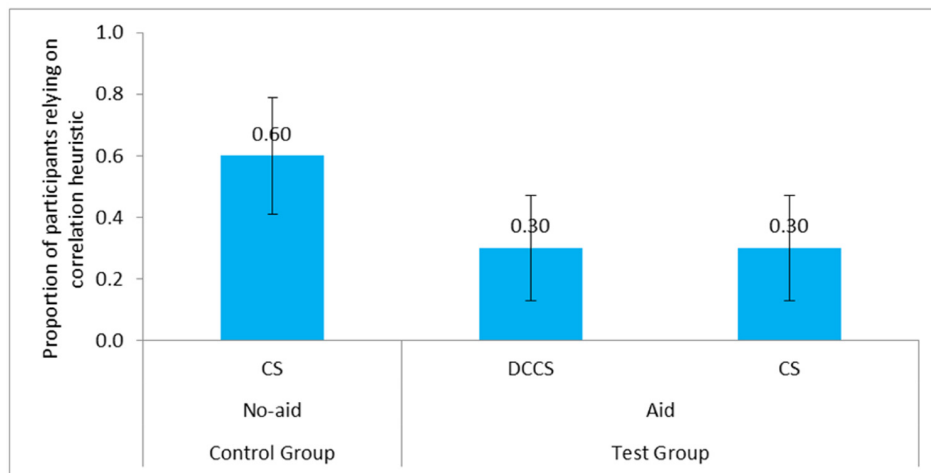
### Violation of Mass Balance

We compared the committing of violation of mass balance between the CS tasks across the Aid and No-aid conditions. **Figure 13** shows the proportion of participants committing violation of mass balance in the Aid and No-aid conditions. Results indicated that violation of mass balance was statistically smaller in the CS task of Aid condition compared to the CS task of No-aid condition [ $0.43 < 0.90$ ,  $\chi^2(1) = 14.70$ ,  $p = 0.00$ ,  $\phi = 0.49$ ]. The proportion of participants committing violation of mass balance in DCCS was close to 0.85. The correlation between the number of times DCCS was used and committing of violation of mass balance in the CS task was small and insignificant ( $r = 0.08$ ,  $p = 0.62$ ).

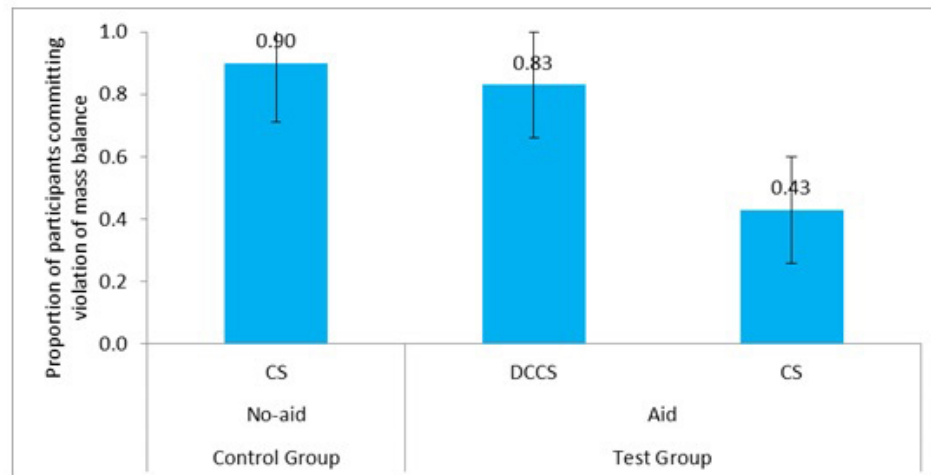
Overall, in agreement with our expectations, the proportion of participants relying on correlation heuristic and committing violation of mass balance was statistically smaller in Aid condition compared to No-aid condition.

## Discussion

Simulations tools may provide effective side-by-side decision aids that enable people to reduce their stock-flow misconceptions. Results revealed that DCCS served as an effective side-by-side decision aid and enabled people to reduce their correlation heuristic and violation of mass balance misconceptions compared to those conditions where DCCS was not present.



**FIGURE 12 |** Proportion of participants relying on correlation heuristic in the Aid and No-aid conditions. The error bars represent 95% confidence interval around the point estimate.



**FIGURE 13 |** Proportion of participants committing violation of mass balance in the Aid and No-Aid conditions. The error bar represents the 95% confidence interval around the point estimate.

One likely reason for the effectiveness of DCCS as a decision aid could be that DCCS enables people to try different scenarios related to how CO<sub>2</sub> emissions and absorptions influence the trajectory of CO<sub>2</sub> concentration (Dörner, 1996; Cronin et al., 2009; Dutt, 2011). Thus, people could use DCCS to try different CO<sub>2</sub> emissions and absorptions values and observe their effect on the resulting CO<sub>2</sub> concentration trajectories. This trial-and-error learning in DCCS is consistent with literature on experienced-based decisions (Jessup et al., 2008; Camilleri and Newell, 2011; Lejarraga and Gonzalez, 2011). For example, according to Jessup et al. (2008), when participants are provided with experiential decision problems, they tend to rely on the experience gained in these problems in making decisions and improve their decision making. Similarly, the experience gained in DCCS enables participants to improve their decision-making in the CS task.

Furthermore, in our results, participants used DCCS between 1 time and 7 times before while attempting the CS task. This use of DCCS agrees with that reported in literature (Hertwig et al., 2004; Cronin et al., 2009). For example, Cronin et al. (2009) gave a stock-flow problem where participants needed to determine the maximum and minimum stock levels across multiple attempts. In each attempt, participants wrote answers to stock questions and they were given feedback on whether their answers were correct or incorrect. According to Cronin et al. (2009), due to the correct-incorrect feedback, more than 70% of the participants were able to answer the stock questions correctly by the fifth attempt (i.e., between one and nine attempts). Similarly, in agreement with our results, Hertwig et al. (2004) have shown that people explore different options presented to them about 7 times before choosing an option for real.



## GENERAL DISCUSSION

In this paper, we started with the general hypothesis that heterogeneity in surface, structure, and problem difficulty in simulation tools as well as the use of the simulation tools as decision aids will be helpful in reducing public stock-flow misconception about Earth's climate. Across the first two experiments, we evaluated how the DCCS enables people to reduce their climate misconceptions because of heterogeneity due to surface and structural features as well as problem difficulty. Also, in a third experiment, we evaluated how DCCS as a side-by-side decision aid helps people to reduce their climate misconceptions. Overall, our results could be explained based upon theoretical arguments concerning the heterogeneity of practice hypothesis (Gonzalez and Madhavan, 2011), procedural reinstatement principle (Schneider et al., 2002; Healy et al., 2005; Young et al., 2011), cognitive load theory (Sweller, 1994; De Jong, 2010), and decisions from experience (Jessup et al., 2008; Camilleri and Newell, 2011; Lejarraga and Gonzalez, 2011).

First, our findings suggest that simulation tools for Earth's climate (like DCCS) are effective in causing learning of both structural features and surface features in problems. In our experiment, people were not given full-information on the formulations connecting emission, absorption, and concentration (Dörner, 1996). These relationships were something that participants had to learn over time while performing in DCCS (Dörner, 1996). Based upon our results, simulation tools like DCCS not only enable people to learn the generality of problems across units and dimensions but also the generality of problems across how inputs and outputs influence the accumulation (Dörner, 1996; Sutton and Barto, 1998; Gonzalez et al., 2003; Dutt and Gonzalez, 2015).

Dutt and Gonzalez (2015) have provided a cognitive account based upon Instance-based Learning Theory (IBLT) on how learning occurs as a dynamic task (like DCCS) due to the focus on process measures and outcome measures. In agreement with Dutt and Gonzalez (2015)'s account, when people come across elements like emission, absorption, and concentration in DCCS, they create instances (or experiences) in their memory. Several experiences get created due to the repeated interaction in DCCS concerning emission, absorption, and concentration values. However, among these instances those instances that allow people to make their CO<sub>2</sub> concentration come closer to the goal are the ones that likely get reinforced over time. While performing the CS task, people retrieve these reinforced instances from memory to make improved decisions. Thus, people likely use their reinforced knowledge acquired in DCCS to draw correct trajectories of emissions and absorptions corresponding to the different concentration curves.

Furthermore, our results revealed that the use of complex curve shapes in simulation tools (i.e., problem difficulty), however, did not help participants to reduce their stock-flow misconceptions. This result could be explained based upon the additional working memory capacity requirements to process complex interaction of different elements like emissions, absorptions, and concentrations (Dörner, 1996). In agreement with cognitive load theory, as our working memory is bounded

(Simon, 1959), people may not be able to process the complex interactions, especially when the concentration curve shapes are complex.

We found that the difficulty hypothesis was unable to account for the findings in the second experiment. One likely reason for this observation could be that the tasks used in our study are different from those that were used for showcasing the difficulty hypothesis (Healy et al., 2005). In literature, the difficulty hypothesis has been showcased using a duration production task in which the dependent measure was reaction time and not the inflow, outflow, and stock. In DCCS, however, the main dependent variables of interest were the inflow, outflow, and stock. Still, another likely reason for the inability of the difficulty hypothesis could be the trajectory of the stock curve used in the difficult condition. It is likely that the stock shapes used in the difficult condition were not difficult enough to cause learning of the underlying relationship between emissions, absorptions, and concentration. Future research should test the learning from complex concentration curves in simulation tools by trying scenarios with concentration curves of different difficulty (Dörner, 1996). Perhaps, stock-flow problems with more challenging CO<sub>2</sub> accumulation curves would be more likely to help reduce correlation heuristic and violation of mass balance misconceptions.

We also found that the stock-flow misconceptions did not reduce when the concentration curve shape in DCCS was simple compared to when people were not exposed to DCCS at all. Thus, overall, this result disagrees with those reported in the first and second experiment, where performance in DCCS caused people to reduce their stock-flow misconceptions compared to conditions where participants were not exposed to DCCS. One likely reason for the disagreement could be the number of repetitions of DCCS given in the second experiment (equal to one) compared to other experiment (multiple). Although we can only speculate currently, but, perhaps, more repetitions of DCCS in simple and difficult conditions could lead people to reduce their stock-flow misconceptions. This hypothesis needs to be tested as part of future research.

Our findings have important implications for real-world climate education as well as climate policymaking. First, as simulation tools like DCCS likely create both surface and structure learning, they are ideal for educating students from kindergarten to standard 12th about stock-flow problems (Gonzalez and Wong, 2012; Meadows et al., 2016). Thus, the use of simulation tools should be encouraged in schools for learning about Earth's climate, especially when students are exposed to concepts like the carbon-cycle and climate change.

Third, the use of simulation tools as decision aids should be encouraged for both climate education and policy analyses. Here, simulation tools can be used as a side-by-side decision aid that provides people the ability to test different hypotheses concerning emissions, absorptions, and concentrations. Also, policymakers could use simulation tools like DCCS for climate policy analyses and to evaluate how different CO<sub>2</sub> emission and absorption trajectories impacts CO<sub>2</sub> concentrations and global temperatures. One expects improved policy analyses with repeated iterations in simulation tools.

The current investigation on the use of simulation tools has revealed promising results. However, there are several research questions to pursue as part of research in the immediate future. Although different structural and surface training was taken into account, comparison with homogenous condition was not made. As part of future research, we would like to compare different structural and surface heterogeneous condition with homogeneous conditions. For example, it would be interesting to analyze how heterogeneity in structure, surface, and problem difficulty interacts with people's science education and other demographic variables. Also, how a group of decision-makers (in contrast to single decision-makers) may improve their correlation heuristic and violation of mass balance misconceptions via simulation tools as well as how these groups show learning of structure, surface, and difficulty? Still, how people who improve their understanding of Earth's climate in problems with a single accumulation (CO<sub>2</sub> concentration) improve their decision-making in problems with two or more accumulations (e.g., CO<sub>2</sub> concentration and global temperatures)? It would be interesting to investigate whether it is people's conscious or unconscious learning that improves due to the use of simulation tools? And, whether people are really learning something about climate change or just learning to use the DCCS tool to complete the CS task?

Prior research has also reported that a part of the stock-flow misconceptions in the CS task could be because of the format of presentation of material concerning emissions, absorptions, and concentration (Fischer et al., 2015). As per Fischer et al. (2015), the use of verbal formats of presentation of stock-flow problems may help reduce some of the stock-flow misconceptions concerning reasoning about stocks. Thus, as part of future research, it would be interesting to test the effectiveness of the heterogeneity in structure, surface, and problem difficulty as well as the extent of learning (conscious or unconscious) in different verbal and non-verbal stock-flow problem formats.

As part of our future work, we would like to answer some of these open-ended questions by involving complex stock-flow problems that vary in their complexity in terms of the number of stock and flows and nature of stock and flows (Frensch and

Funke, 2014). Also, how the increasing complexity of stock-flow problems may interact with the format of presentation of stock-flow problems to influence people's reduction in stock-flow misconceptions (Fischer et al., 2015). Furthermore, one also needs to go deeper to understand the memory processes underlying the learning of structure and surface features in DCCS (Dörner, 1996). Thus, one also needs to evaluate how certain computational models based upon theories of cognition are likely to provide an account of the changes in memory processes in simulation tools (Gonzalez et al., 2003; Gonzalez and Dutt, 2011). We plan to undertake some of these research questions as part of our immediate research on the theme of learning via simulation tools.

## AUTHOR CONTRIBUTIONS

MK was the research lead who designed the experiments and carried out data collection for this work. VD was the principal investigator who served as a constant guiding light for this work.

## ACKNOWLEDGMENTS

The authors would like to thank Indian Institute of Technology Mandi for providing necessary computational for this work. This project was supported by a seed grant (IITM/SG/VD/32) from Indian Institute of Technology Mandi to VD. Also, the authors would like to thank Akanksha Jain, Sushmita Negi, Surendra Singh, and Tushar Galhotra for their help in collecting data across the different experiments reported in this paper.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00299/full#supplementary-material>

**DATA SHEET S1** | Experimental data used in experiments 1, 2 and 3.

## REFERENCES

- Camilleri, A. R., and Newell, B. R. (2011). Description- and experience-based choice: Do equivalent samples equal equivalent choice? *Acta Psychol.* 136, 276–284. doi: 10.1016/j.actpsy.2010.11.007
- Chi, M. T. H., Feltovich, P., and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cogn. Sci.* 5, 121–152. doi: 10.1207/s15516709cog0502\_2
- Cronin, M. A., Gonzalez, C., and Serman, J. D. (2009). Why don't well-educated adults understand accumulation? A challenge to researchers, educators and citizens. *Organ. Behav. Hum. Decis. Process.* 108, 116–130. doi: 10.1016/j.obhdp.2008.03.003
- De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instr. Sci.* 38, 105–134. doi: 10.1007/s11251-009-9110-0
- Dörner, D. (1996). *The Logic of Failure: Recognizing and Avoiding Error in Complex Situations*. New York, NY: Basic Books.
- Dutt, V. (2011). *Why Do We Want To Defer Actions on Climate Change? A Psychological Perspective*. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.
- Dutt, V., and Gonzalez, C. (2012a). Decisions from experience misconceptions about climate change. *J. Environ. Psychol.* 32, 19–29. doi: 10.1016/j.jenvp.2011.10.003
- Dutt, V., and Gonzalez, C. (2012b). Human control of climate change. *Clim. Change* 111, 497–518. doi: 10.1007/s10584-011-0202-x
- Dutt, V., and Gonzalez, C. (2013a). Enabling eco-friendly choices by relying on the proportional-thinking heuristic. *Sustainability* 5, 357–371. doi: 10.3390/su5010357
- Dutt, V., and Gonzalez, C. (2013b). Reducing the linear perception of non-linearity: use of a physical representation. *J. Behav. Decis. Mak.* 26, 51–67. doi: 10.1002/bdm.759
- Dutt, V., and Gonzalez, C. (2015). Accounting for outcome and process measures in dynamic decision-making tasks through model calibration. *J. Dyn. Decis. Mak.* 1, 1–10. doi: 10.11588/jddm.2015.1.17663

- Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G\*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146
- Fischer, H., Degen, C., and Funke, J. (2015). Improving stock-flow reasoning with verbal formats. *Simul. Gaming* 46, 255–269. doi: 10.1177/1046878114565058
- Frensch, P. A., and Funke, J. (eds) (2014). *Complex Problem Solving: The European Perspective*. Hove: Psychology Press.
- Gonzalez, C., and Dutt, V. (2011). A generic dynamic control task for behavioral research and education. *Comput. Hum. Behav.* 27, 1904–1914. doi: 10.1016/j.chb.2011.04.015
- Gonzalez, C., and Wong, H.-Y. (2012). Understanding stocks and flows through analogy. *Syst. Dyn. Rev.* 28, 3–27. doi: 10.1002/sdr.470
- Gonzalez, C., Lerch, F. J., and Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cogn. Sci.* 27, 591–635. doi: 10.1207/s15516709cog2704\_2
- Gonzalez, C., and Madhavan, P. (2011). Diversity during practice enhances detection of novel stimuli. *J. Cogn. Psychol.* 23, 342–350. doi: 10.1080/20445911.2011.507187
- Gonzalez, C., Vanyukov, P., and Martin, M. K. (2005). The use of microworlds to study dynamic decision making. *Comput. Hum. Behav.* 21, 273–286. doi: 10.1016/j.chb.2004.02.014
- Healy, A. F., Wohldmann, E. L., Parker, J. T., and Bourne, L. E. Jr. (2005). Skill training, retention, and transfer: the effects of a concurrent secondary task. *Mem. Cogn.* 33, 1457–1471. doi: 10.3758/BF03193378
- Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychol. Sci.* 15, 534–539. doi: 10.1111/j.0956-7976.2004.00715.x
- Jessup, R. K., Bishara, A. J., and Busemeyer, J. R. (2008). Feedback produces divergence from prospect theory in descriptive choice. *Psychol. Sci.* 19, 1015–1022. doi: 10.1111/j.1467-9280.2008.02193.x
- Lejarraga, T., and Gonzalez, C. (2011). Effects of feedback and complexity on repeated decisions from description. *Organ. Behav. Hum. Decis. Process.* 116, 286–295. doi: 10.1016/j.obhdp.2011.05.001
- McCright, A. M., and Dunlap, R. E. (2011). Cool dudes: the denial of climate change among conservative white males in the United States. *Glob. Environ. Change* 21, 1163–1172. doi: 10.1016/j.gloenvcha.2011.06.003
- Meadows, D., Sweeney, L. B., and Mehers, G. M. (2016). *The Climate Change Playbook*. White River Junction, VT: Chelsea Green Publishing.
- Moxnes, E., and Saysel, A. K. (2009). Misperceptions of global climate change: information policies. *Clim. Change* 93, 15–37. doi: 10.1007/s10584-008-9465-2
- Schneider, V. I., Healy, A. F., and Bourne, L. E. Jr. (2002). What is learned under difficult conditions is hard to forget: contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *J. Mem. Lang.* 46, 419–440. doi: 10.1006/jmla.2001.2813
- Simon, H. (1991). Bounded rationality and organizational learning. *Organ. Sci.* 2, 125–134. doi: 10.1287/orsc.2.1.125
- Simon, H. A. (1959). Theories of decision-making in economics and behavioral science. *Am. Econ. Rev.* 49, 253–283.
- Sterman, J. D. (2008). Risk communication on climate change: mental models and mass balance. *Science* 322, 532–533. doi: 10.1126/science.1162574
- Sterman, J. D. (2011). Communicating climate change risks in a skeptical world. *Clim. Change* 108, 811–826. doi: 10.1007/s10584-011-0189-3
- Sterman, J. D., and Sweeney, L. B. (2002). Cloudy skies: assessing public understanding of global warming. *Syst. Dyn. Rev.* 18, 207–240. doi: 10.1002/sdr.242
- Sterman, J. D., and Sweeney, L. B. (2007). Understanding public complacency about climate change: adults' mental models of climate change violate conservation of matter. *Clim. Change* 80, 213–238. doi: 10.1007/s10584-006-9107-5
- Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. 1(1). Cambridge: MIT press.
- Sweeney, L. B., and Sterman, J. D. (2000). Bathtub dynamics: initial results of a systems thinking inventory. *Syst. Dyn. Rev.* 16, 249–286. doi: 10.1002/sdr.198
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learn. Instr.* 4, 295–312. doi: 10.1016/0959-4752(94)90003-5
- Young, M. D., Healy, A. F., Gonzalez, C., Dutt, V., and Bourne, L. E. Jr. (2011). Effects of training with added difficulties on RADAR detection. *Appl. Cogn. Psychol.* 25, 395–407. doi: 10.1002/acp.1706

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Kumar and Dutt. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# A Cognitive Modeling Approach to Strategy Formation in Dynamic Decision Making

Sabine Prezenski<sup>1\*</sup>, André Brechmann<sup>2\*</sup>, Susann Wolff<sup>2</sup> and Nele Russwinkel<sup>1</sup>

<sup>1</sup> Cognitive Modeling in Dynamic Human-Machine Systems, Department of Psychology and Ergonomics, Technical University Berlin, Berlin, Germany, <sup>2</sup> Special Lab Non-Invasive Brain Imaging, Leibniz Institute for Neurobiology, Magdeburg, Germany

## OPEN ACCESS

### Edited by:

Wolfgang Schoppek,  
University of Bayreuth, Germany

### Reviewed by:

Robert Lawrence West,  
Carleton University, Canada  
David Reitter,  
Pennsylvania State University, United States

### \*Correspondence:

Sabine Prezenski  
sabine.prezenski@tu-berlin.de  
André Brechmann  
brechmann@lin-magdeburg.de

### Specialty section:

This article was submitted to  
Cognitive Science,  
a section of the journal  
Frontiers in Psychology

**Received:** 15 March 2017

**Accepted:** 20 July 2017

**Published:** 04 August 2017

### Citation:

Prezenski S, Brechmann A, Wolff S  
and Russwinkel N (2017) A Cognitive  
Modeling Approach to Strategy  
Formation in Dynamic Decision  
Making. *Front. Psychol.* 8:1335.  
doi: 10.3389/fpsyg.2017.01335

Decision-making is a high-level cognitive process based on cognitive processes like perception, attention, and memory. Real-life situations require series of decisions to be made, with each decision depending on previous feedback from a potentially changing environment. To gain a better understanding of the underlying processes of dynamic decision-making, we applied the method of cognitive modeling on a complex rule-based category learning task. Here, participants first needed to identify the conjunction of two rules that defined a target category and later adapt to a reversal of feedback contingencies. We developed an ACT-R model for the core aspects of this dynamic decision-making task. An important aim of our model was that it provides a general account of how such tasks are solved and, with minor changes, is applicable to other stimulus materials. The model was implemented as a mixture of an exemplar-based and a rule-based approach which incorporates perceptual-motor and metacognitive aspects as well. The model solves the categorization task by first trying out one-feature strategies and then, as a result of repeated negative feedback, switching to two-feature strategies. Overall, this model solves the task in a similar way as participants do, including generally successful initial learning as well as reversal learning after the change of feedback contingencies. Moreover, the fact that not all participants were successful in the two learning phases is also reflected in the modeling data. However, we found a larger variance and a lower overall performance of the modeling data as compared to the human data which may relate to perceptual preferences or additional knowledge and rules applied by the participants. In a next step, these aspects could be implemented in the model for a better overall fit. In view of the large interindividual differences in decision performance between participants, additional information about the underlying cognitive processes from behavioral, psychobiological and neurophysiological data may help to optimize future applications of this model such that it can be transferred to other domains of comparable dynamic decision tasks.

**Keywords:** dynamic decision making, category learning, ACT-R, strategy formation, reversal learning, cognitive modeling, auditory cognition



## INTRODUCTION

Backcountry skiers (and snowboarders) strive for the unique thrill of skiing or snowboarding down powder covered mountains, drawing the first line into freshly fallen snow. Before deciding to go down a particular mountain slope, they check the snowpack, the temperature and wind conditions to avoid setting off an avalanche. Often not a single snow characteristic is crucial but conjunctions of them can change the conditions of safe skiing. The decision to continue on a slope is re-evaluated often, depending on the feedback from the snow (e.g., collapsing snow, snow-brakes vs. nice powder snow) and previous experience.

The described scenario gives a good example of complex cognition. Complex cognition (Knauff and Wolf, 2010) investigates how different mental processes influence action planning, problem solving and decision-making. The term “mental processes in complex cognition” includes not only cognitive but also motivational aspects. Naturalistic decision-making research investigates how decisions are made “in the wild.” Real-life decisions made by people with some kind of expertise are investigated in the context of limited time, conflicting goals, dynamically changing conditions, and information sources of varying reliability.

Such complex situations involve further aspects that cannot all be covered in combination when studying complex cognition. Nevertheless, researchers should aim at describing, understanding and predicting human behavior in its complexity.

A model situated within cognitive architectures can simulate multiple parallel processes, thereby capturing multifaceted psychological phenomena and making predictions, sometimes even for complex tasks. Nevertheless, developing such models requires a stepwise procedure to distinguish different influencing factors. For our skiing example, first a model of the core decision-making process (e.g., based on category learning from snow characteristics and feedback) of a backcountry skier needs to be developed and tested. Afterwards this approach can be extended with modeling approaches of other decision influencing processes (e.g., motivation) to predict decision-making in the wild.

To come closer to the overall goal of understanding cognition as a whole, studying dynamic decision-making with cognitive architectures constitutes a step in the right direction. In dynamic decision-making, decisions are not seen as fixed but can be modified by incoming information. So, not only singular aspects of decision-making are considered, such as attentional influence, but also environmental factors that give feedback about an action or lead to major changes requiring an adaptation to new conditions.

In real-life decisions, however, our future choices and our processing of decision outcomes are influenced by feedback from the environment. This is the interactive view on decision-making, called dynamic decision-making (Gonzalez, 2017), of which the scenario presented above is an example. According to Edwards (1962), three aspects define dynamic decision-making. First, a series of actions are taken over time to achieve a certain goal. Second, the actions depend on each other. Thus, decisions are influenced by earlier actions. Third, and most difficult to

investigate, changes in the environment occur as a result of these actions but also spontaneously (Edwards, 1962). According to Gonzalez (2017), dynamic decision-making is a process where decisions are motivated by goals and external events. They are dependent on previous decisions and outcomes. Thus, decisions are made based on experience and are dependent on feedback. Most of the time, these kinds of decisions are made under time constraints. Therefore, long mental elaborations are not possible. To sum up, dynamic decision-making research investigates a series of decisions which are dependent on previous decisions and are made under time constraints in a changing environment.

Another view on dynamic decision-making as a continuous cycle of mental model updating is introduced by Li and Maani (2011). They describe this process using the CER Cycle. CER stands for Conceptualization–Experimentation–Reflection. Conceptualization is obtaining an understanding of the situation and mentally simulating the outcome of potential decisions and related actions. Thus, the decision maker compares the given situation with related information in his or her mental model and integrates new information obtained from the environment to develop a set of decisions. During experimentation, the decisions and interventions devised from the decision-maker’s mental model are tested in the dynamics of the real world. In the reflection phase, the outcome of the experimentation phase is reflected on, e.g., feedback is processed. If the expected outcome is achieved (e.g., positive feedback), the initial decisions are sustained. If, however, the outcome is unexpected (e.g., negative feedback) or if obtained results differ from the expected outcome, the decision maker updates his or her mental model. To do this, he or she decides for alternative actions such as searching for new sources of information for making better decisions.

These kinds of decision-making procedures have been suggested to share many processes with the procedure of category formation (Seger and Peterson, 2013). Categorization is a mental operation that groups objects based on their similar features. When new categories are formed from a given set of items without explicit instruction, the features distinguishing the different items must first be extracted. Then hypotheses about the relevant features must be formed and tested by making serial decisions.

Category learning experiments in cognitive science often require participants to establish explicit rules that identify the members of a target category. The serial categorization decisions are reinforced by feedback indicating whether a decision was correct or not. The success in such rule-based category learning experiments critically depends on working memory and executive attention (Ashby and Maddox, 2011). The fact that real world decisions critically depend on success and failure in previous trials qualifies category learning as a model for dynamic decision-making.

There are numerous advanced computational models of categorization which explain behavioral performance of subjects in various categorization tasks (e.g., Nosofsky, 1984; Anderson, 1991; Ashby, 1992; Kruschke, 1992; Nosofsky et al., 1994; Erickson and Kruschke, 1998; Love et al., 2004; Sanborn et al., 2010). These competing models differ in their theoretical assumptions (Lewandowsky et al., 2012) and there is currently no

consensus on how different models can be compared and tested against each other (Wills and Pothos, 2012).

Another requirement for dynamic decision-making is the occurrence of changes in the environment. A well-known categorization task using such changes is implemented in the Wisconsin Card Sorting Test (WCST; Berg, 1948). In this test, participants must first select a one-feature rule (color, shape, number of symbols) and are then required to switch to a different one-feature rule. This task tests for the ability to display behavioral flexibility. Another experimental approach to test for behavioral flexibility in humans and animals is reversal learning (e.g., Clark et al., 2004; Jarvers et al., 2016). Here, subjects need to adapt their choice behavior according to reversed reinforcement contingencies.

Thus, category learning experiments with changing rules can serve as suitable paradigms to study dynamic decision-making in the laboratory, albeit with limited complexity as compared to real world scenarios.

The majority of rule-based category learning experiments are simple and only use one relevant stimulus feature specification (e.g., a certain color of the item) as categorization basis. In principle, however, such a restriction is not required and rule-based category learning experiments can become more complex by using conjunction rules. These can still be easily described verbally (e.g., respond A if the stimulus is small on dimension *x* and small on dimension *y*). It has been shown that conjunction rules can be learned (e.g., Salatas and Bourne, 1974) but are much less salient and are not routinely applied (Ashby et al., 1998).

In the following, the main points mentioned above are integrated in our backcountry skiing example: Since feedback from the environment plays a central role in building a correct mental model, feedback in the form of great powder snow indicates that the current strategy is correct. By contrast, negative feedback, for example breaking snow, indicates that one should change the strategy, perhaps search for different feature specifications or even for a different combination of features that might promise a better outcome for skiing. Furthermore, sudden changes of environmental conditions can result in a change of which feature combinations are indicative of a positive outcome. In our example, a change could be a different hill side with more exposure to the sun or a rise in temperature, requiring that other feature combinations should be taken as an indication for a safe descent. There are a lot of possibilities which features and feature combinations could indicate safe or unsafe conditions, making such a task complex.

Thus, to study dynamic decision-making in a category learning experiment requires a task with the above-mentioned characteristics (successive decisions with feedback, multiple feature stimuli, and switching of category assignments). To determine how humans learn feature affiliation in a dynamic environment and to investigate how strategies with rising complexity emerge, a modeling approach addressing these aspects first needs to be developed. If this model is useful and plausible it should match average behavioral data. This is an important milestone toward a more precise model which in turn should predict more detailed empirical data (e.g., individual behavioral or neural data). If this step is achieved, then

models can be used as decision aiding systems at an individual level.

In this paper, we use the behavioral data of an experiment, described in the following, to develop an initial cognitive model as described above. In the experiment, a large variety of multi-feature auditory stimuli were presented to participants in multiple trials. The participants were then required to learn by trial and error which combinations of feature specifications predict a positive or negative outcome. Since perceptual learning of stimulus features is not the focus of our research, we used salient and easy-to-recognize auditory features. To meet all of the above-mentioned criteria for dynamic decision-making, we further introduced a spontaneous change in the environment such that previous decisions on feature combinations suddenly needed to be re-evaluated to obtain positive feedback.

In particular, we would like to demonstrate how different aspects that influence dynamic decision-making can be addressed through a combination of existing and validated cognitive mechanisms within an architecture. These are: learning to distinguish positive and negative feature combinations depending on feedback; successive testing of simple one-feature rules first and switching to more complex two-feature rules later, and using metacognition to re-evaluate feature combinations following environment changes. Other modeling approaches are also able to replicate such data, what distinguishes our approach is that it has a theory grounded interpretation of plausible cognitive mechanisms.

## Why Use Cognitive Modeling?

The method of cognitive modeling forces precision of vague theories. For scientific theories to be precise, these verbal theories should be formally modeled (Dimov et al., 2013). Thus, theories should be constrained by describable processes and scientifically established mechanisms. As Simon and Newell (1971) claim, “the programmability of the theories is a guarantee of their operability and iron-clad insurance against admitting magical entities in the head” (p. 148).

Cognitive models can make predictions of how multiple aspects or variables interact and produce behavior observed in empirical studies. In real-life situations, multiple influences produce behavior. Cognitive models are helpful to understand which interrelated cognitive processes lead to the observed behavioral outcome. Cognitive models can perform the same task as human participants by simulating multiple ongoing cognitive processes. Thereby, models can provide insight into tasks that are too complex to be analyzed by controlled experiments. Nevertheless, studying such a task with participants is mandatory to compare the outcomes of models and participants. However, understanding the process leading to an outcome is more important than perfectly fitting a model to a given set of experimental results. Our goal in this regard is to understand the processes underlying human decision-making, not least to aid humans in becoming better at decision-making (Wolff and Brechmann, 2015).

Predictions made by cognitive models cannot only be compared to average outcome data (such as reaction times, or percentage of correct decisions) but also to process data. Process

data represent patterns of information search, e.g., neural data. In this regard, cognitive models can be informed by EEG and fMRI data to achieve an empirical validation of such processes (Forstmann et al., 2011; Borst and Anderson, 2015).

The development of neurobiologically plausible models is specifically the focus of reinforcement learning (e.g., Sutton and Barto, 1998). The aim of such computational models is to better understand the mechanisms involved on the neural network level as studied using invasive electrophysiological measures in different brain regions in animals (e.g., sensory and motor cortex, basal ganglia, and prefrontal cortex). Such neural network models have recently been applied to learning tasks requiring flexible behavior (e.g., contingency reversal tasks). The reader is referred to a recent paper by Jarvers et al. (2016) that gives an overview of the literature on reversal learning and describes a recurrent neural network model for an auditory category learning task such as the one applied in the current paper. This probabilistic learning model resulted in a good fit to the empirical learning behavior, but does not interpret the cognitive processes that lead to this behavior. It postulates an unspecified metacognitive mechanism that controls the selection of the appropriate strategy. This is where the strength of our approach comes into play; It is specific about the metacognitive mechanisms that drive behavior in such tasks. An example would be processes which assure that after a number of negative results, a change in strategy will be initiated.

To summarize, cognitive modeling is a falsifiable methodology for the study of cognition. In scientific practice, this implies that precise hypotheses are implemented in executable cognitive models. The output of these models (process as well as product) is then compared to empirical data. Fit-Indices such as  $r^2$  and RSME as well as qualitative trends provide information on the predictive power of the cognitive models.

More specific, the central goals of cognitive modeling are to (a) describe (b) predict, and (c) prescribe human behavior (Marewski and Link, 2014). A model that *describes* behavior can replicate the behavior of human participants. If the model, however, reproduces the exact behavior found in the human data, this is an indication of overfitting. In this case, the model has parameters that also fit the noise found in the empirical data. To address such issues of over specified models, it is important to test the model on a new data set and thereby evaluate how well it can *predict* novel data. *Prescribe* means that the model should be a generalizable model so that it can predict behavior in different situations. Moreover, robust models are preferable this implies that the output of the model is not easily influenced by specific parameter settings.

The term cognitive model includes all kinds of models of cognition—from very specific, isolated cognitive aspects only applicable in specific situations to more comprehensive and generalizable ones. The latter candidates are cognitive architectures that consider cognition as a whole. They aim at explaining not only human behavior but also the underlying structures and mechanisms. Cognitive models written on the basis of cognitive architectures therefore generally do not focus on singular cognitive processes, such as some specific learning process. By contrast, interaction of different cognitive processes

and the context of cognitive processes are modeled together. Modeling the relations between different subsystems is especially relevant for applied research questions. The structures and mechanism for this are provided by the cognitive architecture and should be psychologically and neurally plausible (Thomson et al., 2015).

The most commonly used cognitive architectures, such as ACT-R, predict processes at a fine-grain level in the range of 50 ms. These processes can be implemented computationally. However, they are embedded in cognitive theories—this is what distinguishes cognitive models built with cognitive architectures from mathematical models such as neural networks. The latter models formally explain behavior in terms of computational processes. Thus, their explanation of behavior can be seen in terms of computational processes but do not aim at cognitive interpretations (Bowers and Davis, 2012).

## The Cognitive Architecture ACT-R

The cognitive architecture ACT-R (Adaptive Control of Thought—Rational) has been used to successfully model different dynamic decision-making tasks and is a very useful architecture for modeling learning (Anderson, 2007; Gonzalez, 2017). In the following, a technical overview of the main structures and mechanisms that govern cognitive models in ACT-R is given. We will focus only on those aspects that are important to understand our modeling approach. For a more detailed insight into ACT-R, we recommend exploring the ACT-R website<sup>1</sup>.

ACT-R's main goal is to model cognition as a whole using different modules that interact with each other to simulate cognitive processes. These modules communicate via interfaces called buffers. ACT-R is a hybrid architecture, thus symbolic and subsymbolic mechanisms are implemented in the modules of ACT-R.

Our model uses the motor, the declarative, the imaginal, the goal, the aural<sup>2</sup>, and the procedural module. The motor module represents the motor output of ACT-R. The declarative module is the long-term memory of ACT-R in which all information units (chunks) are stored and retrieved. The imaginal module is the working memory of ACT-R in which the current problem state (an intermediate representation important for performing a task) is held and modified. Thus, the imaginal module plays an important role for learning. The goal module holds the control states. These are the subgoals that have to be achieved for the major goal. The aural module is the perceptual module for hearing. The procedural module plays a central role in ACT-R. It is the interface of the other processing units, since it selects production rules (see below) based on the current state of the modules.

Writing a model requires the modeler to specify the symbolic parts of ACT-R. These are (a) the production rules, and (b) the chunks. Chunks are the smallest units of information. All information in ACT-R is stored in chunks. Production rules

<sup>1</sup><http://act-r.psy.cmu.edu/>

<sup>2</sup>Please note that the aural module has two buffers. A tone is first encoded in the aural-location buffer and its content can then be accessed using the aural buffer.

(e.g., productions) consist of a condition and an action part. Productions are selected sequentially, and only one production can be selected at a given time. A production can only be selected if the condition part of the production matches the state of the modules. Then, the action part modifies the chunks in the modules. If more than one production matches the state of the modules then a subsymbolic production selection process chooses which of the matching productions is selected.

A further subsymbolic process in ACT-R is the activation of a chunk. It determines if a chunk can be retrieved from memory and how long this retrieval takes. The past usefulness of a chunk (base-level activation), the chunk's relevance in the current context (associative activation) and a noise parameter sum up to the chunk's activation value. Modifying the subsymbolic mechanisms of ACT-R is also part of the modeling procedure. This can be done using specific parameters—however, most parameters have default values derived from previous studies (Wong et al., 2010) which should be used.

## How Can Decision-Making and Category Learning Be Modeled in ACT-R?

Many different styles for writing models in ACT-R exist (Taatgen et al., 2006). The following modeling approaches have been used for decision-making: (a) strategy or rule-based, (b) exemplar or instance-based, and (c) approaches that mix strategies and exemplars. These approaches will be compared to motivate our chosen modeling approach.

In *strategy or rule-based models*, different problem solving strategies are implemented with different production rules and successful strategies are rewarded. Rule-based theories in category learning postulate that the categorizer must identify the category of an object by testing it against different rules. So, to find a solution for a problem, strategies in the form of rules are used.

*Exemplar or instance-based models* rely on previous experience stored in declarative memory to solve decision-making problems. The content and structure of the exemplars depend on individual framing. It is not a complete representation of the event, but represents the feature specifications the problem solver is focused on, together with experienced feedback. Exemplar theories of category learning postulate that category instances are remembered. To decide if an instance belongs to a category, a new instance is compared to an existing instance. Instance-Based Learning (IBL) builds upon instances in the context of dynamic decision processes and involves learning mechanisms such as recognition-based retrieval. The retrieval of instances depends upon the similarity between the current situation and instances stored in memory. In IBL situations, outcome observations are stored in chunks and retrieved from memory to make decisions. The subsymbolic activation of the retrieved instances determines which instances are likely to be retrieved in a given situation. Instance Based Learning requires some amount of previous learning of relevant instances. Then, decision makers are able to retrieve and generalize from these instances (Gonzalez et al., 2003).

*Mixed approach models* use both rules and instances to solve decision-making problems.

Several authors implemented the described approaches in category learning and decision-making environments. In a strategy-based ACT-R model, Orendain and Wood (2012) implemented different strategies for complex problem solving in a microworld<sup>3</sup> game called “Firechief.” Their model mirrored the behavior of participants in the game. Moreover, different training conditions and resulting behavior of the participants could be modeled. The model performed more or less flexibly, just as the participants, according to different training conditions. This demonstrates that success in strategy learning depends on the succession of stimuli in training conditions. Peebles and Banks (2010) used a strategy-based model of the dynamic-stocks and flows task (DSF). In this task, water level must be held constant but the inflow and outflow of the water changes at varying rates. An ACT-R model of strategies for accomplishing this task was implemented in form of production rules. The model replicated the given data accurately, but was less successful in predicting new data. The authors proposed that by simply extending the model so it contains more strategies and hypotheses, it would be able to predict such new data as well. Thus, specifying adequate rules is crucial for rule-based models.

Gonzalez et al. (2009) compared the performance of two ACT-R models, an instance-based model and a strategy-based model, in a RADAR task. In this task, participants and the model had to visually discriminate moving targets (aircrafts) among moving distractors and then eliminate the targets. Both models achieved about the same overall fit to the participants' data, but IBL performed better in a transfer task.

Lebiere et al. (1998) tested two exemplar models that captured learning during a complex problem-solving task, called the sugar factory (Berry and Broadbent, 1988). The sugar factory task investigates how subjects learn to operate complex systems with an underlying unknown dynamic behavior. The task requires subjects to produce a specific amount of sugar products. Thus, in each trial the workforce needs to be adjusted accordingly. The two exemplar models produced adequate learning behavior similar to that of the subjects. In a subsequent study, Fum and Stocco (2003) investigated how well these original models could predict participants' behavior in case of a much lower target amount of the sugar product than in the original experiment. Furthermore, they investigated if the models could reproduce behavior in case of switching from a high product target amount to a low product target amount and vice versa during the experiment. The performance of the participants increased significantly in the first case. The original IBL models were not able to capture this behavior. The authors therefore developed a rule based model that captured the subjects, switching behavior.

Rutledge-Taylor et al. (2012) compared a rule-based and an exemplar-based model for an intelligence categorization task where learned characteristics had to be studied and assigned. Both models performed equally well in predicting the participants' data. No model was superior to the other.

<sup>3</sup>Microworlds are computer simulations of specific problems. They are applied to study real-world problem solving in dynamic and highly complex settings.



In a different categorization study, Anderson and Betz (2001) studied three category-learning tasks with three different ACT-R models, an exemplar-based model, a rule-based model and a mixed model. The mixed model fitted best, reproducing learning and latency effects found in the empirical data.

In summary, there is no clear evidence that one or the other modeling approach is superior. In their paper, Anderson and Betz (2001) state that the mixed approach is probably the closest to how humans categorize, because the assumption that categorization is either exclusively exemplar-based or exclusively rule-based is probably too close-minded. Furthermore, stimulus succession and adequate rule specification are important for dynamic decision-making and category learning tasks.

In addition, models of complex tasks should incorporate metacognitive processes such as reflecting and evaluating the progress of the selected approach (Roll et al., 2004; Reitter, 2010; Anderson and Fincham, 2014). Reitter's (2010) model of the dynamic stocks and flow tasks investigated how subjects manage competing task strategies. The subject-to-subject analysis of the empirical data showed that participants exhibited sudden marked changes in behavior. Learning mechanisms which are purely subsymbolic cannot explain such behavior, because changes in model behavior would take too long. Furthermore, the strategies of the participants seemed to vary with the complexity of the water flow. Thus, a model of this task must address switches in strategy and not only gradual learning. Reitter (2010) assumes that humans' solutions to real-world problems emerge from a combination of general mechanisms (core learning mechanisms) and decision-making strategies common to many cognitive modeling tasks. His model implements several strategies to deal with the basic control task as well as a mechanism to rank and select those strategies according to their appropriateness in a given situation. This represents the metacognitive aspect of his model.

## Our Aim

Our aim is to develop an ACT-R modeling approach for dynamic decision-making in a category learning task. A suitable task for such a modeling approach needs to fulfill several requirements. First, it should use complex multi-feature stimuli for the model to build categories from combined features. Second, the task needs to provide feedback, thereby allowing the model to learn. Third, changes in the environment should occur during the task forcing the model to act on them by refining once learned category assemblies.

To model performance in such a task, the modeling approach will need to incorporate mechanisms for strategy learning and strategy switching. It should precisely specify how hypotheses about category learning can be implemented with ACT-R. A mixed modeling approach of rules and exemplars should be used since previous work indicates that such models are most suitable for dynamic decision-making tasks. Furthermore, since switches in category assignments as well as monitoring of the learning progress need to be addressed, metacognitive aspects should be incorporated in the modeling approach.

Our modeling approach should provide information on the actual cognitive processes underlying human dynamic

decision-making. Hence, it should be able to predict human behavior and show roughly the same performance effects that can be found in empirical data reflecting decision-making, e.g., response rates. Even more importantly, we aim at developing a general model of dynamic decision-making. For the model to be general (e.g., not fit exclusively to one specific experimental setting or dataset), it needs to be simple. Thus, only few assumptions should be used and unnecessary ones avoided. As a result, the modeling approach should be capable to predict behavior with other stimulus materials and be transferable to other similar tasks.

To summarize the scope of this article, our proposed modeling approach aims to depict the core processes of human decision-making, such as incorporating feedback, strategy updating, and metacognition. Building a model with a cognitive architecture ensures that evaluated cognitive processes are used. The quest is to see whether these cognitive aspects including the processes of the architecture can produce empirical learning behavior:

First, performance improvement through feedback should be included in the model. In the case of feature learning and strategy updating, improvements in one's strategy are only considered in the case of negative feedback (Li and Maani, 2011). If feedback signals a positive decision, people consider their chosen strategy for later use. Thus, people update their mental model during dynamic decision-making only if they receive negative feedback (Li and Maani, 2011). For our feature learning model, this implies that once a successful strategy has been chosen over alternatives, revisions to this strategy will require negative feedback on that strategy rather than positive experience with others, as these are no longer explored.

Second, the model should include transitions from simple to complex strategies. Findings suggest that people initially use simple solutions and then switch to more complex ones (Johansen and Palmeri, 2002). The modeling approach under discussion should be constructed in a similar fashion. In the beginning, it should follow simple one-feature categorization strategies and later switch to more complex two-feature strategies.

Third, the model needs to use metacognitive mechanisms. For example, it needs specifications for which conditions switching from a single-feature strategy to a multi-feature strategy is required. The metacognitive aspects should furthermore reflect previous learning successes. Thus, keeping track of which approaches were helpful and which were not, or of how often a strategy has been successful in the past, should be implemented in the model. Moreover, such mechanisms should ensure that if a strategy was successful in the past and fails for the first time, it is not discarded directly, but tested again. Furthermore, metacognitive mechanisms should not only address the issue of switching from single-feature to multi-feature strategies but also incorporate responses to changes in the environment.

## MATERIALS AND METHODS

In the following, an experiment of dynamic decision-making and our model performing the same task are presented. The model

includes mechanisms to integrate feedback, to switch from simple to complex strategies and to address metacognition. The model was built after the experimental data were obtained.

This section is subdivided in the following manner: First, the participant sample, setup and stimuli of the empirical experiment are described. Then, the modeling approach is explained in detail. Afterwards, the model setup and stimuli are presented. Finally, the analytical methods to evaluate the fit between the model and the empirical results are outlined.

## Experiment Participants

55 subjects participated in the experiment that took place inside a 3 Tesla MR scanner<sup>4</sup> (27 female, 28 male, age range between 21 and 30 years, all right handed, with normal hearing). All subjects gave written informed consent to the study, which was approved by the ethics committee of the University of Magdeburg, Germany.

## Experimental Stimuli

A set of frequency-modulated different tones served as stimuli for the categorization task. The tones differed in duration (short, 400 ms, vs. long, 800 ms), direction of frequency modulation (rising vs. falling), intensity (low intensity, 76–81 dB, vs. high intensity, 86–91 dB), frequency range (five low frequencies, 500–831 Hz, vs. five high frequencies, 1630–2639 Hz), and speed of modulation (slow, 0.25 octaves/s, vs. fast, 0.5 octaves/s), resulting in  $2 \times 2 \times 2 \times 10 \times 2$  (160) different tones. The task relevant stimulus properties were the direction of frequency modulation and sound duration, resulting in four tone categories: short/rising, short/falling, long/rising, and long/falling. For each participant, one of these categories constituted the target sounds (25%), while the other three categories served as non-targets (75%).

As feedback stimuli, we used naturally spoken utterances (e.g., ja, “yes”; nein, “no”) as well as one time-out utterance (zu spät, “too late”) taken from the evaluated prosodic corpus MOTI (Wolff and Brechmann, 2012, 2015).

## Experimental Paradigm

The experiment lasted about 33 min in which a large variety of frequency-modulated tones (see Section Experimental Stimuli above) were presented in 240 trials in pseudo-randomized order and with a jittered inter-trial interval of 6, 8, or 10 s. The participants were instructed to indicate via button-press whether they considered the tone in each trial to be a target (right index finger) or a non-target (right middle finger). They were not informed about the target category but had to learn by trial and error. Correct responses were followed by positive feedback, incorrect responses by negative feedback. If participants failed to respond within 2 s following the onset of the tone, the time-out feedback was presented.

After 120 trials, a break of 20 s was introduced. From the next trial on the contingencies were reversed such that the target stimulus required a push of the right instead of the left button.

The participants were informed in advance about a resting period after finishing the first half of the experiment but they were not told about the contingency reversal.

## Model in Detail

In the following, the model is presented in detail. First, a description of the main declarative representations (chunks) is provided. They reflect strategy representations and metacognitive processes. This is followed by a description of how the model runs through a trial. Finally, the rules that govern strategy learning are summarized.

### Chunks and Production Rules Used in the Model

The chunks implemented in the model are shown in **Figure 1**. “Strategy chunks” hold the strategies in form of examples of feature-value pairs and responses. They are stored in and retrieved from long-term memory (declarative module). The current strategy is held in working memory (imaginal module). Strategy chunks contain the following information about the strategy: which feature(s) and what corresponding value(s) are relevant (e.g., the sound is loud or the sound is loud and its frequency range is high), what the proposed response is (categorization, 1 or 0), and the degree of complexity of the strategy (e.g., one or two-feature strategy). Furthermore, an evaluation mechanism is part of this chunk. This includes noting if a strategy was unsuccessful and keeping track of how often a strategy was successful. This tracking mechanism notices if the first attempt to use this strategy is successful. It then counts the number of successful strategy uses; this explicit count is continued until a certain value is reached. We implemented such a threshold count mechanism to reflect the subjective feeling that a strategy was often useful. We implemented different threshold values for the model. We also differentiated between the threshold for one-feature strategies (first count) and for two-feature strategies (second count). The tracking mechanism can be seen as a metacognitive aspect of our model. Other metacognitive aspects are implemented in the “control chunk” which is kept in the goal buffer of the model. These metacognitive aspects include: first, the level of feature-complexity of the strategy, i.e., if the model attempts to solve the task with a one-feature or with a two-feature strategy; second, whether or not a long-time successful strategy caused an error, this signifies the model’s uncertainty about the accuracy of the current strategy; third, whether changes in the environment occurred that require to renew the search for an adequate strategy.

### Trial Structure

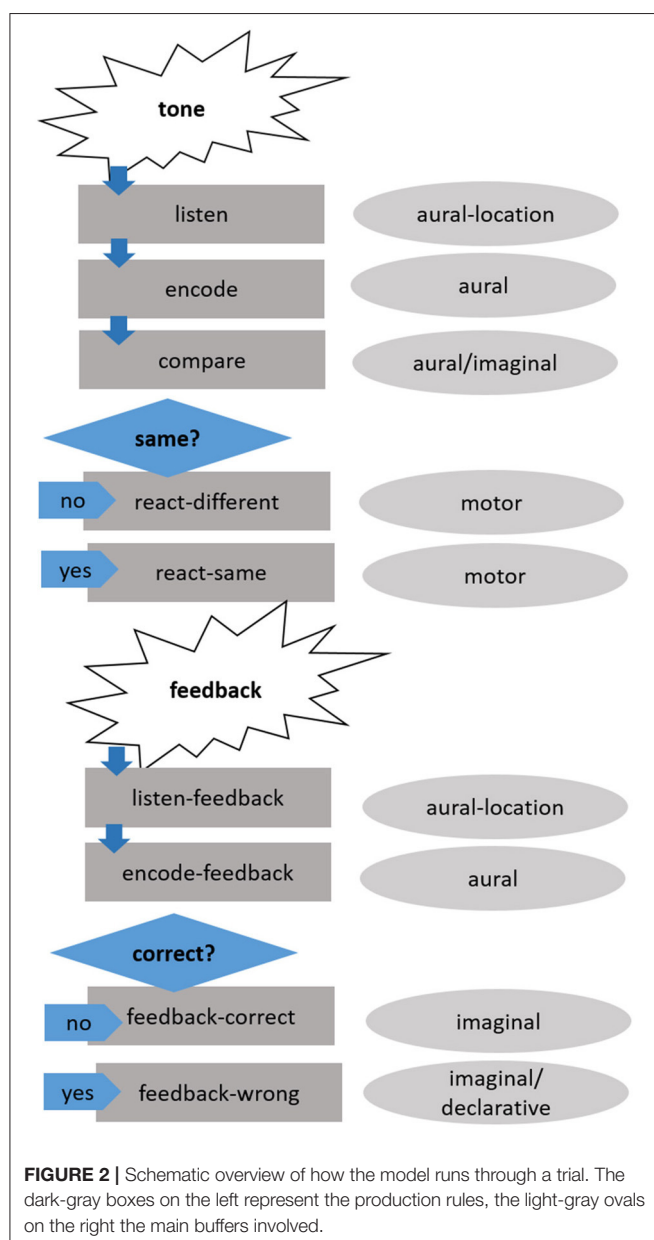
Production rules govern how the model runs through the task. The flow of the model via its production rules is illustrated in **Figure 2**. The following section describes how the model runs through a trial, the specific production rules are noted in parentheses.

A tone is presented to the model and enters the aural-location buffer (*listen*). After the tone has finished, it is encoded in the aural buffer (*encode*). Thus, a chunk with all audio information necessary (duration, direction of pitch change, intensity, and

<sup>4</sup>The experiments were performed inside an MR scanner to study the specific neural correlates of strategy formation which is the subject of another paper.

control chunk		strategy chunk	
complexity	(one or two)	complexity	(one or two)
uncertain	(nil – yes)	1.feature-value-pair	(e.g. duration short)
environment	(nil-change)	2.feature-value-pair	(nil, e.g. volume high)
		response	(0 or 1)
		unsuccessful	(nil-yes)
		first attempt	(nil-yes)
		1.count	(nil, 1,2...threshold)
		2.count	(nil, 1,2...threshold)

**FIGURE 1** | Schematic build-up of the structure of the control and the strategy chunk. Nil indicates that the variable has no value.



frequency range—see Section Modeling Paradigm and Stimuli below) is in the aural buffer and all four characteristics of the tone are accessible to the model. The audio chunk in the aural buffer is then compared to the strategy chunk held in the imaginal buffer (*compare*). If the specific features (e.g., intensity is high) of the strategy chunks are the same as in the audio chunk, the response is according to the strategy proposed by the model (*react-same*), if not, the opposite response is chosen (*react-different*). The presented feedback is listened to and held in the aural-location buffer (*listen-feedback*) and then encoded in the aural buffer (*encode-feedback*). If the feedback is positive, the current strategy is kept in the imaginal buffer and the count-slot is updated (*feedback-correct*). If the feedback is negative, the strategy is updated depending on previous experiences (*feedback-wrong*). Thus, a different strategy chunk is retrieved from declarative memory and copied to the imaginal buffer.

### Finding an Adequate Strategy

All possible strategies are already available in the model's long-term memory. The currently pursued strategy is maintained in working memory and evaluated regarding the feedback. For positive feedback, the strategy is retained and it is counted how often it is successful. If feedback is negative, the strategy is usually altered. The following subsection is a summary of how strategy updating is implemented. For more information see Figure 3.

The model always begins with a one-feature strategy (which strategy it begins with is random) and then switches to another one-feature strategy. The nature of the switch depends on how often a particular strategy was successful. When the model searches for different one-feature strategies, it retrieves only strategies which were not used recently. In case of immediate failure of a one-feature strategy, a different response is used for the feature-value pair. In other cases, the feature-value pair is changed, but the response is retained. If a one-feature strategy has been successful often and then fails once, the strategy is not directly exchanged, but re-evaluated. However, it is also noted that the strategy has caused an error. Two possibilities explain why switches from a one-feature to a two-feature strategy occur: Such a switch can happen either because no one-feature strategy that was not negatively evaluated can be retrieved or because an often successful one-feature strategy failed repeatedly. Switches



### Switching between different one-feature strategies

*If the current strategy is a one-feature strategy and it was not successful at the first attempt then retrieve any other one-feature strategy that has not been evaluated as unsuitable and has not been retrieved recently.*

*If the current strategy is a one-feature strategy and it was successful at the first attempt but has not been successful often then retrieve any other one-feature strategy that has a different feature and has not been evaluated as unsuitable and has not been retrieved recently.*

*If the current strategy is a one-feature strategy and was successful often then keep the strategy, but note that this strategy caused an error.*

### Switching from one-feature to two-feature strategies

*If the current strategy is a one-feature strategy that was successful often but it caused an error before then retrieve a two-feature strategy which includes the current feature value pair and response.*

*If the attempt to retrieve a one-feature strategy has failed then retrieve a two-feature strategy.*

### Switching between different two-feature strategies

*If the current strategy is a two-feature strategy and was not successful at the first attempt then retrieve any other two-feature strategy that has not been evaluated as unsuitable and has not been retrieved recently.*

*If the current strategy is a two-feature strategy and was successful at the first attempt but has not been successful often then retrieve a two-feature strategy which includes the current feature value pair and response but has a different feature and has not been evaluated as unsuitable and has not been retrieved recently.*

*If the current strategy is a two-feature strategy and has been successful often, then keep the strategy but note that this strategy caused an error.*

### Detecting a change

*If the current strategy is a two-feature strategy and it caused an error before then retrieve a different two-feature strategy and note that something in the environment has changed and evaluate the strategy as unsuitable for this change.*

### Searching for a strategy after a detected change

*If the current strategy is a two-feature strategy and a change in the environment is detected and the first attempt to use this strategy was not successful then retrieve a different two-feature strategy that has not been evaluated as unsuitable for this change.*

*If the current strategy is a two-feature strategy and a change in the environment is detected and the strategy was successful more than once but not often, then keep one feature and the response and retrieve a strategy that has not been evaluated unsuitable for this change.*

**FIGURE 3 |** Rules governing when and to what degree the strategies are changed after negative feedback is received.

within the two-feature strategy are modeled the following way: If a two-feature strategy was unsuccessful at the first attempt, any other two-feature strategy is used (which one exactly is random). If a two-feature strategy was initially successful and then fails, then a new strategy which retains one of the feature-value pairs and the response will be selected. This strategy only differs in the other feature-value pair. When the environment changes, a

previously often successful two-feature strategy (and also a one-feature strategy) will fail. Then a retrieval of another two-feature strategy is attempted. If at the time the environment changes, the model has not found a successful two-feature strategy, it will continue looking for a useful two-feature strategy, and thus not notice the change.

## Modeling Paradigm and Stimuli

The following section briefly describes how the experiment was implemented for the model. This includes a short overview of how the stimulus presentation was modified for the model.

The task of the participants was implemented for the model in ACT-R 7.3 with some minor modifications. The same four pseudo-randomizations used for the participants were also used for the model. Thus, 25% of the stimuli were target stimuli. A trial began with a tone, which lasted for 400 ms. To model the two stimulus durations, we used two different features in the new-other-sound command. As soon as the model responded via button press, auditory feedback was presented. Overall, a trial lasted for a randomized period of 6, 8, or 10 s, similar to the original experiment. There was no break for the model after 120 trials, but the targets switched after 120 trials, too.

Instead of employing all 160 different tones, sixteen different tones were presented to the model. Each of the tones is a composition of four characteristics of the four binary features: duration (long vs. short), direction of frequency modulation (rising vs. falling), intensity (low intensity, vs. high intensity), and frequency range (low vs. high). Only binary features were used for the model because the perceptual difference between the two classes of each selected feature was high, except for speed of modulation, which was therefore not implemented in the model. For the participants, more feature variations were used to ensure categorical decisions and to prevent them from memorizing individual tone-feedback pairs. This is not an issue for the model, since no mechanism allowing such memorizing was implemented. As for the participants, auditory feedback was presented to the model.

The modeling approach is a mixed modeling approach, the strategies are encoded as instances, but which instance is retrieved is mainly governed by rules.

To test if the model is a generalizable model, different variations were implemented. The learning curves found in the empirical data should still be found under different plausible parameter settings. However, specific parameter settings should influence the predictive quality of the model. The approach typically chosen by cognitive modelers is to search for specific parameter settings that result in an optimal fit and then report this fit. The objective behind such an approach is to show that the model resembles the ongoing cognitive processes in humans. We have chosen a different approach. Our objective is to show that our modeling approach can map the general behavior such as learning and reversal learning as well as variance found in the data. By varying parameter settings, we want to optimize the fit of the model and examine the robustness of the model mechanisms to parameter variations.

Regarding the choice of varying parameters, we use an extended parameter term which includes not only subsymbolic



ACT-R parameters (which are typically regarded as parameters) but also certain (production) rules (Stewart and West, 2010). In the case of this model, productions that control the tracking mechanism of successful strategies are varied. The tracking mechanism keeps track of how often a strategy is successful. However, the model does not increase the count throughout the entire experiment. After it reaches a threshold, a successful strategy is marked as “successful often.” Thereafter, it is not discharged directly in case of negative feedback but instead reevaluated. So, to answer the question what the most suitable values for the threshold of the first and second count are, these values were varied. Another implemented model assumption is that this threshold is different for single-feature vs. two-feature strategies. We assumed that the threshold for two-feature strategies should be double the value for one-feature strategies, as if the model was counting for each feature separately. The first count was varied for three, four and five and the second count for six, eight, and ten.

Besides the parameters that control the tracking mechanism, we also investigated a parameter-controlled memory mechanism. The latter controls for how long the model can remember if it had already used a previous strategy. This is the *declarative-finst-span*<sup>5</sup> parameter of ACT-R. We assumed that participants remember which strategy they previously used for around 10 trials back. We therefore tested two different values (80 and 100 s) for this parameter, determining whether the model can remember if this chunk has been retrieved in the last 80 (or 100) s. The combination of the declarative-finst-span (80, 100), three values for the first count (3, 4, 5) and three values for the second count (6, 8, 10) resulted in 18 modeling versions (see Table 1).

## Analyses

Each of the models was run 160 times, 40 times for each pseudo-randomized order, using ACT-R 7.3. The data were preprocessed with custom Lisp files and then analyzed with Microsoft Excel.

The model data and the empirical data were divided into 12 blocks, with 20 trials per block. The average proportion of correct responses and the standard deviation per block was computed for the experiment as well as for each of the 18 models.

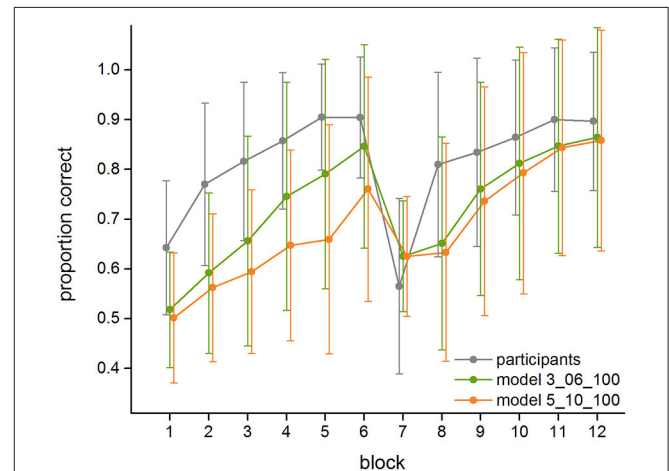
One aim of this study was to predict average learning curves of the participants. Thus, the proportion of correct responses of the participants was compared to the proportion of correct responses of each of the models. Visual graphs comparing the modeled to the empirical data were analyzed with regard to increases and decreases in correct responses.

As an indication of relative fit, the correlation coefficient ( $r$ ) and the determination coefficient ( $r^2$ ) were computed. They represent how well trends in the empirical data are captured by the model.

As an indication of absolute fit, the root-mean-square error (RMSE) was calculated. RMSE represents how accurately the

**TABLE 1 |** Resulting modeling versions from combining the different parameter settings for the first and second count and the declarative-finst-span.

		First count 3	First count 4	First count 5
:declarative-finst-span 80	Second count 6	3_06_080	4_06_080	5_06_080
	Second count 8	3_08_080	4_08_080	5_08_080
	Second count 10	3_10_080	4_10_080	5_10_080
:declarative-finst-span 100	Second count 6	3_06_100	4_06_100	5_06_100
	Second count 8	3_08_100	4_08_100	5_08_100
	Second count 10	3_10_100	4_10_100	5_10_100



**FIGURE 4 |** Average performance and standard deviations of the human participants, the best fitting model (3\_06\_100), and the worst fitting model (5\_10\_100) in the 12 blocks of the experiment.

model predicts the empirical data. RMSE is interpreted as the standard deviation of the variance of the empirical data that is not explained by the model.

To compare the participant-based variance found in the empirical data with the variance produced by the 160 individual model runs, a Levene's test (a robust test for testing the equality of variances) was calculated for each block of the experiment.

## RESULTS

In the following sections, the empirical data, the modeled learning curves, and the results regarding the general fit of the different model versions to the data are presented.

### Empirical Learning Curves

The descriptive analysis of the empirical data (see Figure 4 and Table 2) shows that on average, in the first block the participants respond correctly in 64.3% ( $\pm 13.5\%$ ) of the trials. The response rate of the participants increases until the sixth block to 90.4% ( $\pm 12.2\%$ ) of correct trials. In the seventh block, the block in which targets and non-targets switch, it drops to 56.5% ( $\pm 17.7\%$ ) of correct trials. It then increases again and reaches 81.0% ( $\pm 18.5\%$ ) of correct trials in the eighth block and 89.7% ( $\pm 13.9\%$ ) of correct trials in the last block. Across all 12 blocks, the standard

<sup>5</sup>The declarative-finst-span parameter controls how long a finst (fingers of instantiation) can indicate that a chunk was recently retrieved. The number of items and the time for which an item can be tagged as attended is limited. These attentional markers are based on the work of Zenon Pylyshyn.

**TABLE 2 |** Average proportion of correct responses and standard deviations (in %) of the participants and the 18 versions of the model in the 12 blocks of the experiment.

	1	2	3	4	5	6	7	8	9	10	11	12
Participants	64.3 ± 13.5	77.0 ± 16.4	81.6 ± 15.9	85.7 ± 13.7	90.5 ± 10.7	90.4 ± 12.2	56.5 ± 17.7	81.0 ± 18.5	83.4 ± 18.9	86.4 ± 15.6	90.0 ± 14.4	89.7 ± 13.9
<b>MODELS</b>												
3_06_080	51.1 ± 13.9	59.0 ± 15.5	67.3 ± 20.9	73.9 ± 22.3	75.9 ± 24.5	82.8 ± 21.1	62.9 ± 11.8	64.4 ± 20.2	71.3 ± 22.5	73.9 ± 25.9	81.0 ± 22.3	83.8 ± 23.6
3_06_100	51.8 ± 11.6	59.2 ± 16.1	65.6 ± 21.0	74.6 ± 22.9	79.1 ± 23.0	84.7 ± 20.4	62.5 ± 11.1	65.1 ± 21.4	76.1 ± 21.4	81.2 ± 23.4	84.7 ± 21.5	86.4 ± 22.1
3_08_080	50.1 ± 11.7	56.0 ± 16.0	62.0 ± 17.9	68.5 ± 20.7	72.3 ± 23.8	81.5 ± 20.9	62.8 ± 11.4	66.7 ± 20.4	75.6 ± 22.4	77.6 ± 25.6	83.7 ± 22.1	87.0 ± 21.3
3_08_100	50.0 ± 11.5	57.9 ± 14.9	66.7 ± 21.1	73.8 ± 22.2	77.2 ± 23.9	86.5 ± 19.8	61.9 ± 12.3	61.6 ± 21.2	73.8 ± 24.0	79.9 ± 22.7	82.8 ± 23.3	87.2 ± 20.7
3_10_080	51.3 ± 11.9	54.5 ± 16.8	66.9 ± 20.6	71.8 ± 22.4	75.9 ± 24.2	82.7 ± 22.5	62.6 ± 11.7	63.6 ± 19.4	70.3 ± 22.7	75.6 ± 23.6	81.1 ± 21.6	83.6 ± 22.9
3_10_100	51.6 ± 12.5	60.6 ± 17.1	66.3 ± 21.0	73.5 ± 22.8	74.1 ± 25.4	83.8 ± 20.6	63.0 ± 11.4	65.3 ± 21.1	74.7 ± 23.1	79.8 ± 23.8	84.2 ± 22.7	87.7 ± 21.1
4_06_080	50.5 ± 12.3	58.5 ± 16.0	65.6 ± 20.3	71.7 ± 20.8	72.2 ± 24.7	81.6 ± 20.9	61.6 ± 12.2	64.9 ± 20.9	73.5 ± 24.1	76.7 ± 24.4	82.6 ± 22.6	83.7 ± 23.9
4_06_100	52.3 ± 11.8	58.3 ± 15.4	65.5 ± 20.9	71.4 ± 21.8	74.5 ± 23.4	81.9 ± 21.1	62.2 ± 11.8	63.9 ± 20.9	74.0 ± 22.4	76.0 ± 25.5	84.2 ± 21.6	85.7 ± 21.9
4_08_080	50.3 ± 12.6	56.2 ± 14.0	61.0 ± 17.7	70.9 ± 21.5	72.9 ± 24.4	78.4 ± 23.4	64.0 ± 11.1	63.1 ± 21.6	75.3 ± 22.8	77.7 ± 25.1	84.3 ± 20.7	85.0 ± 23.3
4_08_100	51.1 ± 12.1	57.8 ± 15.1	63.3 ± 20.0	68.7 ± 20.8	70.1 ± 24.5	78.7 ± 23.4	62.8 ± 12.7	66.2 ± 20.5	74.8 ± 23.1	77.6 ± 25.2	83.3 ± 21.9	85.9 ± 22.0
4_10_080	49.5 ± 11.3	58.9 ± 15.5	63.0 ± 20.1	69.9 ± 21.1	70.2 ± 25.6	80.0 ± 21.8	63.3 ± 11.3	61.3 ± 20.1	70.0 ± 23.7	73.2 ± 24.7	81.7 ± 21.8	81.9 ± 24.8
4_10_100	51.7 ± 12.5	56.8 ± 16.1	64.5 ± 16.9	68.3 ± 22.2	71.9 ± 24.3	81.3 ± 22.4	63.3 ± 11.7	64.2 ± 20.7	72.6 ± 22.7	79.0 ± 23.2	83.5 ± 22.4	86.3 ± 22.0
5_06_080	51.3 ± 12.2	56.5 ± 14.3	61.0 ± 16.2	66.8 ± 20.1	69.1 ± 23.3	78.8 ± 21.9	61.4 ± 11.7	64.1 ± 20.5	72.8 ± 22.9	74.9 ± 25.0	83.2 ± 21.9	85.1 ± 22.8
5_06_100	53.0 ± 11.6	58.5 ± 16.5	61.3 ± 18.8	65.1 ± 21.2	65.9 ± 23.1	76.2 ± 20.7	59.8 ± 11.9	64.1 ± 19.7	72.0 ± 23.0	74.7 ± 24.3	81.3 ± 21.6	85.9 ± 21.2
5_08_080	50.2 ± 11.2	55.3 ± 16.2	58.6 ± 16.0	64.3 ± 19.5	65.3 ± 22.6	72.6 ± 22.2	58.8 ± 12.2	62.6 ± 19.6	72.7 ± 21.3	75.6 ± 23.1	82.6 ± 21.9	86.6 ± 20.9
5_08_100	49.1 ± 12.0	55.0 ± 14.1	60.7 ± 18.2	67.4 ± 20.2	67.0 ± 23.4	74.1 ± 22.6	61.2 ± 12.0	63.8 ± 21.2	73.1 ± 22.4	73.5 ± 24.1	81.5 ± 21.8	83.4 ± 23.3
5_10_080	50.3 ± 12.5	56.4 ± 14.8	60.8 ± 18.9	68.1 ± 19.7	69.3 ± 23.6	76.5 ± 22.7	62.3 ± 13.4	65.5 ± 21.9	74.9 ± 23.0	76.4 ± 25.7	84.5 ± 20.3	85.5 ± 22.0
5_10_100	50.2 ± 13.1	56.2 ± 14.9	59.4 ± 16.4	64.7 ± 19.2	65.9 ± 23.0	76.0 ± 22.5	62.5 ± 12.0	63.3 ± 21.9	73.6 ± 23.0	79.3 ± 24.3	84.3 ± 21.7	85.8 ± 22.2
<b>MODELS OVERALL</b>												
MEAN	50.8	57.3	63.3	69.6	71.6	79.9	62.2	64.1	73.4	76.8	83.0	85.4
MIN	49.1	54.5	58.6	64.3	65.3	72.6	58.8	61.3	70.0	73.2	81.0	81.9
MAX	53.0	60.6	67.3	74.6	79.1	86.5	64.0	66.7	76.1	81.2	84.7	87.7

deviation of the empirical data ranges from 10.7% minimum to 18.9% maximum, with an average standard deviation of 15.1%. The standard deviation of the participants derives from the fact that different participants showed different learning curves, and not all participants reported to have found the correct strategy in a post interview. Correspondingly, eleven participants (20.0%) showed a performance below 85% by the end of the first part of the experiment (Block 6), and 12 participants (21.8%) stayed below 85% correct responses at the end of the second part (Block 12).

## Modeled Learning Curves

**Figure 4** further shows the means and standard deviations of the proportion of correct responses of the best (3\_06\_100) and worst fitting (5\_10\_100) model (see below, Section Model Fit). In addition, **Table 2** lists the model performance means and standard deviations for each of the twelve blocks for all 18 models, and **Figure 5** shows the learning curves of all 18 models.

Both the best and the worst fitting model (as do all others) capture the overall shape of the learning curve found in the data. They both show an increase in the learning rate in the first six blocks. Similarly, all models show a drop in performance in the seventh block, which is followed by another increase in performance. Even in the best fitting 3\_06\_100 model, however, the proportion of correct responses is underestimated by the model, especially in the first blocks. Also, the participants show a more severe setback after the switch but then recover faster, while the model takes longer until its performance increases again. Nevertheless, for the best fitting model, the modeled data are always within the range of the standard deviation of the empirical data.

As **Table 2** shows, each of the models shows a large degree of variance across its 160 runs. The standard deviation averaged across all 12 blocks ranges from 18.9 to 20.4%, depending on the model's parameter settings. For the best-fitting model, the standard deviation in the individual blocks ranges from 11.6 to 23.4% and is significantly larger than the standard deviation

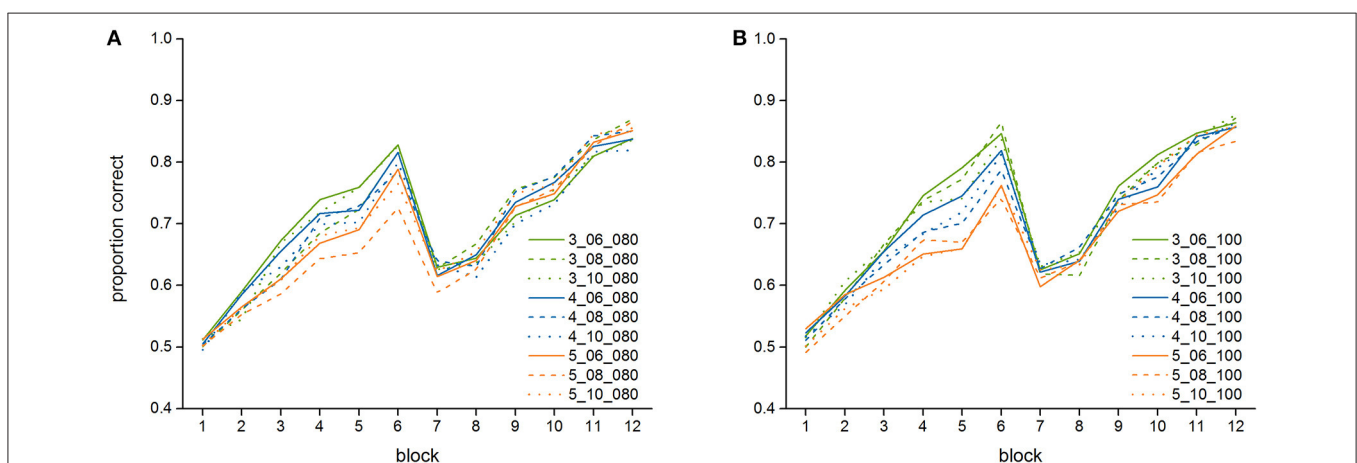
found in the empirical data, except for the first two blocks of the experiment and the first two blocks after the switch (for all blocks except Block 1, 2, 7, and 8: all  $F_s > 6.79$ , all  $p_s < 0.010$ ). This high variation of the individual model runs indicates that the same underlying rule-set with the same parameter settings can still result in very different learning curves, depending on which exact strategies are chosen at each point when a new strategy is selected (e.g., initial strategy, alteration of one-feature strategy, alteration of two-feature strategy). Furthermore, similarly to the non-learners among the participants described above (see Section Empirical Learning Curves), not every model run was successful, resulting (for the best fitting model) in a performance below 85% in 35.6% of the runs for Block 6 and in 30.0% of the runs for Block 12.

## Model Fit

The average correlation of the model and the empirical data is 0.754. Between 43.9% and 67.1% of the variance in the data is explained by the different models. The average standard deviation of the unexplained variance is 0.136. All  $r$ ,  $r^2$ , and RMSE values for the 18 model versions are presented in **Table 3**.

As **Table 3** and **Figure 5** show, the model shows relative robustness to the influence of varying parameter settings. For the first count, a lower value is somewhat better for the fit—there is a stronger increase in the first part of the experiment (until Block 6) for a lower than for a higher first count value. For the second count, a lower value results in a better fit as well. The influence of the declarative-first-span parameter on the fit-indices is very small, resulting in a slightly better fit either for a declarative-first-span of 80 s or of 100 s, depending on the settings of first and second count.

The best fit in terms of correlation was achieved for the model with the declarative-first-span value set to 100 (i.e., the model was able to remember if it had already used a previous strategy for 100 s), a first count of three (i.e., a one-feature strategy needed to be successful at least three times to be considered as “often successful”) and a second count of six (i.e., a two-feature strategy



**FIGURE 5 |** Average performance of the 18 versions of the model in the 12 blocks of the experiment, **(A)** models with a declarative-first-span of 80 s, **(B)** models with a declarative-first-span of 100 s.

**TABLE 3 |** Values of  $r$ ,  $r^2$ , and RMSE of the 18 versions of the model.

	$r$	$r^2$	RMSE
3_06_080	0.812	0.659	0.124
3_06_100	0.820	0.672	0.109
3_08_080	0.745	0.555	0.134
3_08_100	0.803	0.645	0.119
3_10_080	0.785	0.616	0.132
3_10_100	0.798	0.636	0.114
4_06_080	0.805	0.649	0.128
4_06_100	0.794	0.631	0.124
4_08_080	0.726	0.527	0.138
4_08_100	0.743	0.552	0.135
4_10_080	0.745	0.555	0.146
4_10_100	0.741	0.549	0.133
5_06_080	0.733	0.537	0.146
5_06_100	0.722	0.521	0.152
5_08_080	0.697	0.485	0.164
5_08_100	0.718	0.516	0.158
5_10_080	0.721	0.520	0.144
5_10_100	0.663	0.439	0.156
<b>OVERALL</b>			
MEAN	0.754	0.570	0.136
MIN	0.663	0.439	0.109
MAX	0.820	0.671	0.164

needed to be successful at least six times to be considered as “often successful”). The worst fit was observed for the model with the declarative-first-span value set to 100, a first count of five and a second count of ten.

The RMSE varies from a minimum of 0.106 (3\_06\_100) to a maximum of 0.164 (5\_08\_100). Thus, the model with a first count of three, a second count of six and a declarative-first-span set to 100 performs best, both in terms of correlation ( $r$ ) and absolute prediction (RMSE).

## Summary

In general, the models predict the data well. The modeled learning curves resemble the form of the average empirical learning curve, with an increase in the first half of the experiment, a short decrease at the beginning of the second half, followed by another increase in performance. The correlation indices of the best fitting model show a good fit, with 67.2% of the variance of the data being explained by the model with a declarative-first span of 100 s, a first count threshold of three and second count threshold of six. Note that this is also the model with the closest absolute fit (RMSE is 0.109).

However, in absolute percentages of correct responses, all of the models perform below the participants in all blocks (except Block 7). Also, the models show greater overall variance than the empirical data. Furthermore, the models are initially less affected by the switch in strategies but take longer to “recover” from the switch in strategies.

In summary, the model replicates the average learning curves and large parts of the variance. It does so with a limited set of

rules and the given exemplars, covering learning and relearning processes which take place in dynamic environments. Moreover, we found differences in model fit depending on the exact specification of the parameters, with the best fit if the model remembers previously employed strategies for 100 s, marks a one-feature strategy as “often successful” after three successful uses and a two-feature after six successful uses. However, all of the 18 different parameter settings we tested resembled the main course of the empirical data, thereby indicating that the mechanisms of the model are robust to parameter variations.

## DISCUSSION

The discussion covers three main chapters. First, the fit of the model is discussed and suggestions for possible improvements are given. Second, the broader implications of our approach are elaborated. Finally, future work is outlined.

### Discussion of the Modeling Approach

Our modeling account covers relevant behavioral data of a dynamic decision-making task in which category learning is required. To solve the task, two features have to be combined, and the relevant feature combination needs to be learned by trial and error using feedback. The model uses feedback from the environment to find correct categories and to enable a switch in the assignment of response buttons to the target and non-target categories. Metacognition is built into the model via processes that govern under what conditions strategic changes, such as transitions from one-feature to two-feature strategies, occur.

Overall, the fit indices indicate that this model solves the task in a similar way as participants do. This includes successful initial learning as well as the successful learning of the reversal of category assignment. Moreover, the observation was made that not all participants are able to solve the task, and the same is observed in the behavior of the modeling approach. Thus, the model is able to generate output data that, on a phenomenological level, resemble those of subjects performing a dynamic decision-making task that includes complex rule learning and reversal processes. Although the overall learning trends found in the data can be replicated well with the general rules implemented in our model, there are two limitations: The variance of the model is larger than that of the participants, and the overall performance of the model is lower than the performance of the participants.

It is likely that the participants have a different and perhaps more specific set of rules than the model. For example, the participants were told which of the two keys to press for the target sound. However, it is unclear if they used this knowledge to solve the task. To keep the model simple, it was not given this extra information, so there was no meaning assigned to the buttons. This is one possibility to explain the model's lower performance, especially in the first block. Another example for more task specific rules used by the participants compared to the model is that the four different features of the stimuli may not be equally salient to the subjects, which may have led to a higher performance compared to the model. For example,



it is conceivable that the target-feature direction of frequency modulation (up vs. down) was chosen earlier in the experiment than the non-target feature frequency range, while the model treated all features equally to keep the model as simple as possible. Finally, after the change of the button press rule, some participants might have followed a rule which states to press the opposite key if a strategy was correct for many times and then suddenly is not, instead of trying out a different one- or two-feature strategy, whereas the model went the latter way.

Adding such additional rules and premises to the model would possibly reduce the discrepancy between the performance of the model and the behavioral data. However, the aim of this paper was to develop a modeling approach that incorporates general processes important for all kinds of dynamic decision-making. This implies using only assumptions that are absolutely essential (meta-cognition, switching from one-feature to two-feature strategies, learning via feedback) and keeping the model as simple as possible in other regards. As a consequence, adding extra rules would not produce a better general model of dynamic decision-making, but would only lead to a better fit of the model for a specific experiment while making it prone to overfitting. As mentioned earlier, good descriptive models capture the behavioral data as closely as possible and therefore always aim at maximizing the fit to the data they describe. Good predictive models, on the other hand, should be generalizable to also predict behavior in different, but structurally similar situations and not just for one specific situation with one set of subjects. In our view, this constitutes a more desirable quest with more potential to understand the underlying processes of human dynamic decision-making. This is supported by Gigerenzer and Brighton (2009), who argue that models that focus on the core aspects of decision-making, e.g., considering only few aspects, are closer to how humans make decisions. They also argue that such simplified assumptions make decisions more efficient and also more effective (Gigerenzer and Brighton, 2009).

As stated earlier, one way to model dynamic decision-making in ACT-R using only few assumptions is instance based learning (IBL). This approach uses situation-outcome pairs and subsymbolic strengthening mechanisms for learning. However, IBL is insufficient to model tasks which involve switches in the environment (Fum and Stocco, 2003). Such tasks require adding explicit switching rules. Besides these rules, our task needed mechanisms that control when to switch from simple one-feature strategies to more complex strategies. Since meta-cognitive reflections are not part of IBL, we used a mixed modeling approach which incorporates explicit rules and metacognitive reflection. IBL is insofar part of our approach as the strategies are encoded as situation-outcome pairs and subsymbolic strengthening mechanisms of ACT-R are utilized.

To evaluate if our modeling approach of strategy formation and rule switching is in line with how participants perform in such tasks, data reflecting learning success need to be considered. Such data are the learning curves reported in this paper. We believe that an IBL model alone cannot produce the strong increase in performance after the environmental change in the empirical data.

For a further understanding of complex decision-making, other behavioral data, such as reaction times, could also be modeled. However, not all processes that probably have an impact on reaction time are part of our general modeling approach. This is especially the case for modeling detailed aspects of auditory encoding with ACT-R; for example, the precise encoding of the auditory events can be expected to comprise a different gain in reaction time for short compared with longer tones. However, our modeling approach is expandable, allowing the incorporation of other cognitive processes such as more specific auditory encoding or attention. This extensibility is one of the strengths of cognitive architectures and is particularly relevant for naturalistic decision-making, where many additional processes eventually need to be considered.

## Scope of the Model

A formal model was built with ACT-R, it specifies the assumptions of dynamic decision-making in category learning. This model was tested on empirical data and showed similar learning behavior. Assumptions about how dynamic decisions in category learning occur, e.g., by learning from feedback and switching from simple to more complex strategies, and metacognitive mechanisms were modeled together. ACT-R aims at modeling cognition as a whole, thus addressing different cognitive processes simultaneously, an important aspect for modeling realistic cognitive tasks. Moreover, the model is flexible. Thus, the model chooses from the available strategies according to previous experience and random influences.

Our modeling approach is simple in the sense that it comprises only few plausible assumptions, does not rely on extra parameters and is nevertheless flexible enough to cope with dynamically changing environments.

To test the predictive power of the model, it needs to be further tested and compared to new empirical data that are obtained using slightly different task settings. Our aim was to develop a first model of dynamic decision-making in category learning. Thus, relevant cognitive processes that occur between stimulus presentation and the actual choice response are included in the model. Furthermore, we wanted to show how a series of decisions emerge in the pursuit of an ultimate goal. Thus, as a first step we needed a decision task that shows characteristics similar to natural dynamic settings. Such aspects include complex multi-feature stimuli, feedback from the environment, and changing conditions. Since explicit hints on category membership are usually not present in non-experimental situations, it is furthermore reasonable to use a task without explicit instructions regarding which features (or stimuli) attention should be focused on. The downside of using unspecific instructions as done in our study is that from the behavioral data, it will remain unclear how exactly individual participants process such a task, since aspects such as which exact rules are followed or which features are considered at the beginning of a task, are uncertain.

As a next step we aim at modeling and predicting the dynamic decision-making course of individual participants. In general, a big advantage of cognitive modeling approaches is that they can predict ongoing cognitive processes at any point in time. To

evaluate the validity of such predictions, different approaches can be followed.

One approach to constructing models in accordance with the cognitive processes of participants is the train-to-constrain paradigm (Dimov et al., 2013). This paradigm requires instructing participants in a detailed step-by-step procedure on how to apply specific strategies in decision tasks. This approach gives the modeler insight into the strategies that participants are using at a given time point. This again can be used to constrain ACT-R models in the implementation of these strategies. In future studies, we plan to adopt this paradigm by (a) instructing the participants and (b) adjusting our model accordingly. To ensure that the train-to-constrain paradigm was successfully implemented, self-reports of the participants should be used.

Another approach is to conduct interviews while the participant is performing the task. To confirm the model's predictions about the prospective behavior of participants, subjects of future empirical studies should thus be asked about their decisions during the course of the experiment. The first few participant decisions can be expected to be strongly influenced by random aspects (e.g., which feature is attended to first), but after some trials, the modeling approach should be able to predict the next steps of the participants. Thus, it should allow precise predictions of the subsequent cognitive processes. To make such predictions, a revised model would need to use the first couple of trials as information about the strategy an individual participant initially follows.

In a further step, the exact cognitive processes proposed by the model should be tested on an individual level on more fine grain data (e.g., fMRI) and then be readjusted accordingly. Currently, different methods to map cognitive models to finer grain data such as fMRI or EEG data have been proposed (Borst and Anderson, 2015; Borst et al., 2015; Prezenski and Russwinkel, 2016a). These methods are currently investigated and have been applied for basic research questions. Nevertheless, mapping cognitive models to neuronal data is a challenge. More research is needed especially for applied tasks. To supplement neuronal data, additional behavioral data, such as button press dynamics (e.g., intensity of button press), can be added as an immediate measurement of how certain an individual participant is about a decision (Kohrs et al., 2014).

Besides using cognitive models to predict individual behavior, we aim to develop more general cognitive mechanisms to model learning, relearning and metacognition that are valid in a broad range of situations. To test the applicability of our modeling approach in a broader context and different situations, variations of the experiment should be tested with different tasks and materials. For example, the model proposed here should be able to predict data from categorization experiments using visual stimuli such as different types of lamps (Zeller and Schmid, 2016) with some modifications to the sensory processing of our model. Furthermore, the model should be capable of predicting data from different types of categorization tasks, for example a task using a different number of categorization features, more switches or different sequences. Such a task would be a predictive challenge for our model; if it succeeds, it can be considered as a predictive model.

The developed general mechanisms can also be used in sensemaking tasks. Such tasks require “an active process to construct a meaningful and functional representation of some aspects of the world” (Lebiere et al., 2013, p. 1). Sensemaking is an act of finding and interpreting relevant facts amongst the sea of incoming information, including hypothesis updating. Performance in our task comes close to how people make sense in the real world because it involves a large number of different stimuli, each carrying different specifications of various features. Thus, “making sense of the stimuli” requires the participants to validate each stimulus in a categorical manner and use the extracted stimulus category in combination with the selected button-press and the feedback that follows as information for future decisions.

To conclude, such a cognitive model which includes general mechanism for learning, relearning and metacognition can prove extremely useful for predicting individual behavior in a broad range of tasks. However, uncertainty remains regarding whether this captures the actual processes of human cognition. This is not only due to the fact that human behavior is subject to manifold random influences, but also to the limitation that a model always corresponds to a reduced representation of reality. The modeler decides which aspects of reality are characterized in the model. Marewski and Mehlhorn (2011) tested different modeling approaches for the same decision-making task. While they found that their models differed in terms of how well they predicted the data, they ultimately could not show that the best fitting model definitely resembles the cognitive processes of humans. To our knowledge, no scientific method is ever able to answer how human cognition definitely works. In general, models can only be compared in terms of their predictive quality (e.g., explained variance, number of free parameters, generalizability). Which model ultimately corresponds to human reality, on the other hand, cannot be ascertained.

## Outlook

One reason for modeling in cognitive architectures is to implement cognitive mechanisms in support systems for complex scenarios. Such support systems mainly use machine learning algorithms. Unfortunately, those algorithms depend on many trials to learn from before they succeed in categorization or in learning in general. Cognitive architecture inspired approaches, on the other hand, can also learn from few samples. In addition, approaches that rely on cognitive architectures are informed models that provide information about the processes involved and the reasons that lead to success and failure.

Cognitive models can be applied to a variety of real-world tasks, for example to predict usability in smartphone interaction (Prezenski and Russwinkel, 2014, 2016b), air traffic control (Taatgen, 2001; Smieszek et al., 2015), or driving behavior (Salvucci, 2006). Moreover, cognitive modeling approaches can also be used in microworld scenarios (Halbrügge, 2010; Peebles and Banks, 2010; Reitter, 2010). Not only can microworld scenarios simulate the complexity of the real world, they also have the advantage of being able to control variables. This implies that specific variations can be induced to test the theoretical approach or model in question (as demonstrated in Russwinkel et al., 2011).

Many applied cognitive models are quite specific task models. Our model, in contrast, aims at capturing core mechanisms found in a variety of real world tasks. As a consequence, it has the potential to be applied in many domains. So, our model of dynamic decision-making in a category learning task makes predictions about the cognitive state of humans during such a task. This involves predictions about strategies (e.g., one-feature or two-feature strategies), conceptual understanding (e.g., assumptions about relevant feature combinations) and metacognitive aspects (e.g., information on the success of the decision maker's current assumption), all of which are aspects of cognition in a multitude of tasks and application domains.

Our general modeling approach therefore has the potential to support users in many domains and in the long run could be used to aid decision-making. For this, the decisions of individual users during the course of a task could be compared to the cognitive processes currently active in the model. If for example a user sticks to a one-feature strategy for too long or switches rules in an unsystematic manner, a system could provide the user with a supportive hint. Other than regular assistant systems, such a support system based on our model would simulate the cognitive state of the user. For example, this online support system would be able to predict the influence of reoccurring negative feedback on the user, e.g., leading him to attempt a strategy change. If, however, the negative feedback was caused by an external source such as a technical connection error, opting for the strategy change would result in frustration of the user. The proposed support system would be able to intervene here. Depending on the internal state of the user, the support system would consider what kind of information is most supportive or if giving no information at all is appropriate (e.g., in case of mental overload of the user). As long as no support is needed, systems like this would silently follow the decisions made by a person.

Moreover, if the goal of the user is known, and the decisions made by the user have been followed by the system, it would be possible to predict the user's next decisions and also to evaluate whether those decisions are still reasonable to reach the goal. Many avalanches have been caused by repeated wrong

decisions by backcountry skiers stuck in their wrong idea about a situation (Atkins, 2000). A support system that is able to understand when and why a person is making unreasonable decisions in safety critical situations would also be able to present the right information to overcome the misunderstanding. A technical support system for backcountry skiers would need information about current avalanche danger, potential safe routes and other factors. Such information is already provided by smartphone applications that use GPS in combination with weather forecasts and slope-steepness measures. In the future, when this information is made available to a cognitive model-based companion system that predicts the decisions of the users, it could potentially aid backcountry skiers. Cognitive model-based support systems designed in a similar manner could equally well be employed in other safety-critical domains, as well as to assist cyclist, drivers or pilots.

## AUTHOR CONTRIBUTIONS

AB and SW designed the auditory category learning experiment. SW performed human experiments and analyzed the data. SP and NR designed the ACT-R modeling. SP implemented ACT-R modeling and analyzed the data. SP, SW, and AB prepared figures. SP, NR, and AB drafted manuscript. SP, NR, AB, and SW edited, revised and approved the manuscript.

## FUNDING

This work was done within the Transregional Collaborative Research Centre SFB/TRR 62 "A Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG) and supported by funding by the BCP program.

## ACKNOWLEDGMENTS

We thank Monika Dobrowolny and Jörg Stadler for support in data acquisition within the Combinatorial Neuroimaging Core Facility (CNI) of the Leibniz Institute for Neurobiology.

## REFERENCES

- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychol. Rev.* 98, 409–429. doi: 10.1037/0033-295X.98.3.409
- Anderson, J. R. (2007). *How can the Human Mind Occur in the Physical Universe?* New York, NY: Oxford University Press.
- Anderson, J. R., and Betz, J. (2001). A hybrid model of categorization. *Psychon. Bull. Rev.* 8, 629–647. doi: 10.3758/BF03196200
- Anderson, J. R., and Fincham, J. M. (2014). Extending problem-solving procedures through reflection. *Cogn. Psychol.* 74, 1–34. doi: 10.1016/j.cogpsych.2014.06.002
- Ashby, F. G. (1992). "Multidimensional models of categorization," in *Multidimensional Models of Perception and Cognition*, ed F. G. Ashby (Hillsdale, NJ: Erlbaum), 449–483.
- Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., and Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychol. Rev.* 105, 442–481.
- Ashby, F. G., and Maddox, W. T. (2011). Human category learning 2.0. *Ann. N. Y. Acad. Sci.* 1224, 147–161. doi: 10.1111/j.1749-6632.2010.05874.x
- Atkins, D. (2000). "Human factors in avalanche accidents," *Proceedings. Int'l Snow Science Workshop* (Big Sky: MT), 46–51.
- Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. *J. Gen. Psychol.* 39, 15–22. doi: 10.1080/00221309.1948.9918159
- Berry, D. C., and Broadbent, D. E. (1988). Interactive tasks and the implicit-explicit distinction. *Br. J. Psychol.* 79, 251–272. doi: 10.1111/j.2044-8295.1988.tb02286.x
- Borst, J. P., and Anderson, J. R. (2015). "Using the ACT-R Cognitive Architecture in Combination with fMRI data," in *An Introduction to Model-Based Cognitive Neuroscience*, eds B. U. Forstmann, and E.-J. Wagenmakers (New York, NY: Springer), 339–352.
- Borst, J. P., Nijboer, M., Taatgen, N. A., Van Rijn, H., and Anderson, J. R. (2015). Using data-driven model-brain mappings to constrain formal models of cognition. *PLoS ONE* 10:e0119673. doi: 10.1371/journal.pone.0119673
- Bowers, J. S., and Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychol. Bull.* 138, 389–414. doi: 10.1037/a0026450
- Clark, L., Cools, R., and Robbins, T. W. (2004). The neuropsychology of ventral prefrontal cortex: decision-making and reversal learning. *Brain Cogn.* 55, 41–53. doi: 10.1016/S0278-2626(03)00284-7



- Dimov, C. M., Marewski, J. N., and Schooler, L. J. (2013). "Constraining ACT-R models of decision strategies: an experimental paradigm," in *Cooperative Minds: Social Interaction and Group Dynamics. Proceedings of the 35th Annual Conference of the Cognitive Science Society*, eds M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth (Austin, TX: Cognitive Science Society), 2201–2206.
- Edwards, W. (1962). Dynamic decision theory and probabilistic information processing. *Hum. Factors* 4, 59–73. doi: 10.1177/001872086200400201
- Erickson, M. A., and Kruschke, J. K. (1998). Rules and exemplars in category learning. *J. Exp. Psychol.* 127, 107–140. doi: 10.1037/0096-3445.127.2.107
- Forstmann, B. U., Wagenmakers, E.-J., Eichele, T., Brown, S., and Serences, J. T. (2011). Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract? *Trends Cogn. Sci.* 15, 272–279. doi: 10.1016/j.tics.2011.04.002
- Fum, D., and Stocco, A. (2003). "Instance vs. rule-based learning in controlling a dynamic system," in *Proceedings of the Fifth International Conference on Cognitive Modelling* (Bamberg: Universitaets-Verlag Bamberg), 105–110.
- Gigerenzer, G., and Brighton, H. (2009). Homo heuristics: why biased minds make better inferences. *Top. Cogn. Sci.* 1, 107–143. doi: 10.1111/j.1756-8765.2008.01006.x
- Gonzalez, C. (2017). "Decision making: a cognitive science perspective," in *The Oxford Handbook of Cognitive Science (Vol. 1)*, ed S. E. F. Chipman (Oxford: Oxford University Press), 249–263.
- Gonzalez, C., Dutt, V., Healy, A. F., Young, M. D., and Bourne, L. E. Jr. (2009). "Comparison of instance and strategy models in ACT-R," in *Proceedings of the 9th International Conference on Cognitive Modeling—ICCM2009*, eds A. Howes, D. Peebles, and R. Cooper (Manchester).
- Gonzalez, C., Lerch, J. F., and Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cogn. Sci.* 27, 591–635. doi: 10.1207/s15516709cog2704\_2
- Halbrügge, M. (2010). Keep it simple - A case study of model development in the context of the Dynamic Stocks and Flows (DSF) task. *J. Artif. Gen. Intell.* 2, 38–51. doi: 10.2478/v10229-011-0008-2
- Jarvers, C., Brosch, T., Brechmann, A., Woldeit, M. L., Schulz, A. L., Ohl, F. W., et al. (2016). Reversal learning in humans and gerbils: dynamic control network facilitates learning. *Front. Neurosci.* 10:535. doi: 10.3389/fnins.2016.00535
- Johansen, M. K., and Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cogn. Psychol.* 45, 482–553. doi: 10.1016/S0010-0285(02)00505-4
- Knauff, M., and Wolf, A. G. (2010). Complex cognition: the science of human reasoning, problem-solving, and decision-making. *Cogn. Process.* 11, 99–102. doi: 10.1007/s10339-010-0362-z
- Kohrs, C., Hrabal, D., Angenstein, N., and Brechmann, A. (2014). Delayed system response times affect immediate physiology and the dynamics of subsequent button press behavior. *Psychophysiology* 51, 1178–1184. doi: 10.1111/psyp.12253
- Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychol. Rev.* 99, 22–44. doi: 10.1037/0033-295X.99.1.22
- Lebiere, C., Pirolli, P., Thomson, R., Paik, J., Rutledge-Taylor, M., Staszewski, J., et al. (2013). A functional model of sensemaking in a neurocognitive architecture. *Comput. Intell. Neurosci.* 2013, 1–29. doi: 10.1155/2013/921695
- Lebiere, C., Wallach, D., and Taatgen, N. A. (1998). "Implicit and explicit learning in ACT-R," in *Proceedings of the Second European Conference on Cognitive Modelling*, eds F. Ritter and R. Young (Nottingham: Nottingham University Press), 183–193.
- Lewandowsky, S., Palmeri, T. J., and Waldmann, M. R. (2012). Introduction to the special section on theory and data in categorization: integrating computational, behavioral, and cognitive neuroscience approaches. *J. Exp. Psychol. Learn. Mem. Cogn.* 38, 803–806. doi: 10.1037/a0028943
- Li, A., and Maani, K. (2011). "Dynamic decision-making, learning and mental models," in *Proceedings of the 29th International Conference of the System Dynamics Society* (Washington, DC), 1–21.
- Love, B. C., Medin, D. L., and Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychol. Rev.* 111, 309–332. doi: 10.1037/0033-295X.111.2.309
- Marewski, J. N., and Link, D. (2014). Strategy selection: an introduction to the modeling challenge. *Wiley Interdiscipl. Rev.* 5, 39–59. doi: 10.1002/wcs.1265
- Marewski, J. N., and Mehlhorn, K. (2011). Using the ACT-R architecture to specify 39 quantitative process models of decision making. *Judgm. Decis. Mak.* 6, 439–519. doi: 10.101112/jdm101112.pdf
- Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *J. Exp. Psychol.* 10, 104–114. doi: 10.1037/0278-7393.10.1.104
- Nosofsky, R. M., Palmeri, T., and McKinley, S. (1994). Rule-plus-exception model of classification learning. *Psychol. Rev.* 101, 53–79. doi: 10.1037/0033-295X.101.1.53
- Orendain, A. D. O., and Wood, S. (2012). "An account of cognitive flexibility and inflexibility for a complex dynamic task," in *Proceedings of the 11th International Conference on Cognitive Modeling* (Berlin), 49–54.
- Peebles, D., and Banks, A. P. (2010). Modelling dynamic decision making with the ACT-R cognitive architecture. *J. Artif. Gen. Intell.* 2, 52–68. doi: 10.2478/v10229-011-0009-1
- Prezenski, S., and Russwinkel, N. (2014). Combining cognitive ACT-R models with usability testing reveals users mental model while shopping with a smartphone application. *Int. J. Adv. Intell. Syst.* 7, 700–715.
- Prezenski, S., and Russwinkel, N. (2016a). "A proposed method of matching ACT-R and EEG-Data," in *Proceedings of the 14th International Conference on Cognitive Modeling*, eds D. Reitter, and F. E. Ritter (University Park, PA: Penn State), 249–251.
- Prezenski, S., and Russwinkel, N. (2016b). "Towards a general model of repeated app usage," in *Proceedings of the 14th International Conference on Cognitive Modeling*, eds D. Reitter and F. E. Ritter (University Park, PA: Penn State), 201–207.
- Reitter, D. (2010). Metacognition and multiple strategies in a cognitive model of online control. *J. Artif. Gen. Intell.* 2, 20–37. doi: 10.2478/v10229-011-0007-3
- Roll, I., Baker, R. S., Aleven, V., and Koedinger, K. R. (2004). "A metacognitive ACT-R model of students' learning strategies in Intelligent Tutoring Systems," in *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, eds J. C. Lester, R. M. Vicari, and F. Paragacu (Maceió), 854–856.
- Russwinkel, N., Urbas, L., and Thüring, M. (2011). Predicting temporal errors in complex task environments: a computational and experimental approach. *Cogn. Syst. Res.* 12, 336–354. doi: 10.1016/j.cogsys.2010.09.003
- Rutledge-Taylor, M., Lebiere, C., Thomson, R., Staszewski, J., and Anderson, J. R. (2012). "A comparison of rule-based versus exemplar-based categorization using the ACT-R architecture," in *Proceedings of 21st Annual Conference on Behavior Representation in Modeling and Simulation, BRIMS 2012* (Amelia Island, FL), 44–50.
- Salatas, H., and Bourne, L. E. Jr. (1974). Learning conceptual rules: III. Processes contributing to rule difficulty. *Mem. Cogn.* 2, 549–553. doi: 10.3758/BF03196919
- Salvucci, D. D. (2006). Modeling driver behavior in a cognitive architecture. *Hum. Factors* 48, 362–380. doi: 10.1518/00187200677724417
- Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychol. Rev.* 117, 1144–1167. doi: 10.1037/a0020511
- Seger, C. A., and Peterson, E. J. (2013). Categorization = decision making + generalization. *Neurosci. Biobehav. Rev.* 37, 1187–1200. doi: 10.1016/j.neubiorev.2013.03.015
- Simon, H. A., and Newell, A. (1971). Human problem solving: the state of the theory in 1970. *Am. Psychol.* 26, 145–159. doi: 10.1037/h0030806
- Smieszek, H., Joeres, F., and Russwinkel, N. (2015). "Workload of airport tower controllers: empirical validation of a macro-cognitive model," in *Online Journal Kognitive Systeme*, eds D. Soeffker, and A. Kluge (Duisburg: DuEPublico). doi: 10.17185/duepublico/37699
- Stewart, T. C., and West, R. (2010). Testing for equivalence: a methodology for computational cognitive modeling. *J. Artif. Gen. Intell.* 2, 69–87. doi: 10.2478/v10229-011-0010-8
- Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. London: MIT Press.
- Taatgen, N. A. (2001). "A model of individual differences in learning Air Traffic Control," in *Proceedings of the fourth International Conference on Cognitive Modeling*, eds E. M. Altmann, A. Cleeremans, C. D. Schunn, and W. D. Gray (Mahwah, NJ: Erlbaum), 211–216.
- Taatgen, N. A., Lebiere, C., and Anderson, J. R. (2006). "Modeling Paradigms in ACT-R," in *Cognition and Multi-Agent Interaction: From Cognitive Modeling*



- to *Social Simulation*, ed R. Sun (Cambridge: Cambridge University Press), 29–52.
- Thomson, R., Lebiere, C., Anderson, J. R., and Staszewski, J. (2015). A general instance-based learning framework for studying intuitive decision-making in a cognitive architecture. *J. Appl. Res. Mem. Cogn.* 4, 180–190. doi: 10.1016/j.jarmac.2014.06.002
- Wills, A. J., and Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychol. Bull.* 138, 102–125. doi: 10.1037/a0025715
- Wolff, S., and Brechmann, A. (2012). “MOTI: a motivational prosody corpus for speech-based tutorial systems,” in *Proceedings of the 10th ITG Conference on Speech Communication* (Berlin: IEEE), 1–4.
- Wolff, S., and Brechmann, A. (2015). Carrot and Stick 2.0: the benefits of natural and motivational prosody in computer-assisted learning. *Comput. Hum. Behav.* 43, 76–84. doi: 10.1016/j.chb.2014.10.015
- Wong, T. J., Cokely, E. T., and Schooler, L. J. (2010). “An online database of ACT-R parameters: Towards a transparent community-based approach to model development,” in *Proceedings of the 10th International Conference on Cognitive Modeling*, eds D. D. Salvucci, and G. Gunzelmann (Philadelphia, PA: Drexel University), 282–286.
- Zeller, C., and Schmid, U. (2016). “Rule learning from incremental presentation of training examples: Reanalysis of a categorization experiment,” in *Proceedings of the 13th Biannual Conference of the German Cognitive Science Society*. (Bremen), 39–42.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Prezenski, Brechmann, Wolff and Russwinkel. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Advantages of publishing in Frontiers



## OPEN ACCESS

Articles are free to read  
for greatest visibility  
and readership



## FAST PUBLICATION

Around 90 days  
from submission  
to decision



## HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,  
and constructive  
peer-review



## TRANSPARENT PEER-REVIEW

Editors and reviewers  
acknowledged by name  
on published articles

## Frontiers

Avenue du Tribunal-Fédéral 34  
1005 Lausanne | Switzerland

**Visit us:** [www.frontiersin.org](http://www.frontiersin.org)

**Contact us:** [info@frontiersin.org](mailto:info@frontiersin.org) | +41 21 510 17 00



## REPRODUCIBILITY OF RESEARCH

Support open data  
and methods to enhance  
research reproducibility



## DIGITAL PUBLISHING

Articles designed  
for optimal readership  
across devices



## FOLLOW US

@frontiersin



## IMPACT METRICS

Advanced article metrics  
track visibility across  
digital media



## EXTENSIVE PROMOTION

Marketing  
and promotion  
of impactful research



## LOOP RESEARCH NETWORK

Our network  
increases your  
article's readership