Education Tech Research Dev
DOI 10.1007/s11423-013-9301-x

DEVELOPMENT ARTICLE

Computer-based assessment of Complex Problem Solving: concept, implementation, and application

Samuel Greiff · Sascha Wüstenberg · Daniel V. Holt · Frank Goldhammer · Joachim Funke

Abstract Complex Problem Solving (CPS) skills are essential to successfully deal with environments that change dynamically and involve a large number of interconnected and partially unknown causal influences. The increasing importance of such skills in the 21st century requires appropriate assessment and intervention methods, which in turn rely on adequate item construction, delivery, and scoring. The lack of assessment tools, however, has slowed down research on and understanding of CPS. This paper first presents the MicroDYN framework for assessing CPS, which is based on linear structural equation systems with input and output variables and opaque relations among them. Second, a versatile assessment platform, the CBA Item Builder, which allows the authoring, delivery, and scoring of CPS tasks for scientific and educational purposes is introduced. Third, we demonstrate the potential of such a tool for research by reporting an experimental study illustrating the effect of domain specific content knowledge on performance in CPS tasks both on an overall performance and on a process level. The importance of accessible and versatile technical platforms not only for assessment and research but also for intervention and learning are discussed with a particular focus on educational contexts.

Keywords Complex Problem Solving · MicroDYN · Technical platform · Process data · CBA Item Builder · Content knowledge

S. Greiff (✉) · S. Wüstenberg
EMACS Unit, University of Luxembourg, 6, rue Richard Coudenhove Kalergi,
1359 Luxembourg-Kirchberg, Luxembourg
e-mail: samuel.greiff@uni.lu

S. Wüstenberg
e-mail: sascha.wuestenberg@uni.lu

D. V. Holt · J. Funke
University of Heidelberg, Heidelberg, Germany

F. Goldhammer
German Institute for International Educational Research (DIPF) and Centre for International Student Assessment (ZIB), Frankfurt, Germany
e-mail: goldhammer@dipf.de

Published online: 26 April 2013

Introduction

Analyzing data gathered over several decades by the U.S. Department of Labor, Autor et al. (2003) found that almost everywhere, be it in industries, professions, or education, computerization and automation was associated with a reduction of routine tasks and with an increase of tasks that are interactive, complex, and dynamic (Frensch and Funke 1995). This development has been observed in a number of countries and appears particularly distinctive in service-oriented Western economies (e.g., Autor and Dorn 2009; Spitz-Oener 2006). These changes in the job environment are directly associated with changes in skills required to be successful at the work place. For example, there is an increase in the demand for Complex Problem Solving (CPS) skills, the ability to successfully negotiate complex and unknown situations by applying behavioral patterns beyond routine solutions (Funke 2001; Holyoak 1985). According to Mayer (2003), a problem arises when the difference between a given state and a desired goal state cannot be resolved by applying existing solution patterns (Funke 2003). The subsequent process of problem solving has been defined as successfully transforming the given state into the goal state by applying different problem solving skills (Mayer and Wittrock 2006). Specifically, solving complex problems involves initiating interactions with a dynamically changing and previously unknown environment to gather knowledge about this environment and subsequently control it in order to reach a desired goal state (Raven 2000). Buchner (1995, p. 14) defines CPS as:

> The successful interaction with task environments that are dynamic (i.e., change as a function of user's intervention and/or as a function of time) and in which some, if not all, of the environment's regularities can only be revealed by successful exploration and integration of the information gained in that process.

CPS can be decomposed into two overarching processes: acquiring new knowledge ("knowledge acquisition") and applying this knowledge ("knowledge application") while interacting with a dynamically changing system (Funke 2001; Mayer and Wittrock 2006). Some research suggests to further separate these two processes into sub-processes such as generating and reducing information during knowledge acquisition or systematically intervening during knowledge application (Dörner 1986; Fischer et al. 2012). One of the strategies relevant for knowledge acquisition in CPS is the Vary-One-Thing-at-a-Time strategy (VOTAT; Tschirgi 1980), which describes the systematic variation of single elements in a problem situation to identify isolated effects of these elements (Klahr 2000; Vollmeyer et al. 1996). For instance, Chen and Klahr (1999) demonstrated the importance of systematic variation and VOTAT in solving complex science problems, but this strategy can be applied to a range of other problem domains.

Measures of CPS have been shown to be conceptually and empirically different from intelligence (Gonzalez et al. 2005; Greiff and Fischer 2013) and incrementally predict relevant outcomes such as academic achievement (Wüstenberg et al. 2012) or supervisor ratings (Danner et al. 2011) beyond intelligence. Furthermore, it has been argued that CPS is a suitable candidate for assessing relevant skills in interactive environments (Kröner et al. 2005). In fact, CPS is a major domain in the international large-scale Programme for International Student Assessment (PISA) 2012 run by the Organization for Economic Cooperation and Development (OECD). The OECD has set up a strategy to assist its member countries in exploiting the potential of problem solving skills and transforming them into educational and job outcomes (OECD 2010, 2012). Ideally, the process of acquiring CPS skills continues over the entire lifetime (as assumed in the emerging field of lifelong learning; e.g., Smith and Reio 2006), but formal education is a crucial phase for the development of such skills with corresponding implications for educational systems (OECD 2012).

The assessment of CPS relies strongly on the availability of flexible tools and platforms for authoring and computer-based delivery of complex simulations as assessment vehicles for CPS. Interestingly, Williamson et al. (2006) identify dynamically changing situations—a main feature of CPS—as the specific application of computer-based assessment with the highest added value, while noting that this potential is currently insufficiently exploited. More than 20 years ago, Bunderson et al. (1989) anticipated a swift and general shift towards computer delivered testing. They expected technical platform availability to be a minor issue. However, from the current perspective neither conceptual issues nor psychometric constraints are obstructing the comprehensive advance towards computer-based assessment of CPS, whereas authoring, delivery, and scoring still remain major challenges. A widespread application of computer-based assessment necessitates versatile and easy-to-use assessment technology as a link between theoretically motivated assessment of CPS skills and technical realization.

This paper focuses on computer-based assessment of CPS skills employing the MicroDYN framework (Greiff et al. 2012), which was used in PISA 2012 as conceptualization of CPS (OECD 2010). MicroDYN is a generic theoretical and assessment framework for developing CPS tasks, which is based on theories from cognitive psychology, relies on computer-based testing, and has been empirically validated in various studies (e.g., Greiff et al. 2012; Wüstenberg et al. 2012).

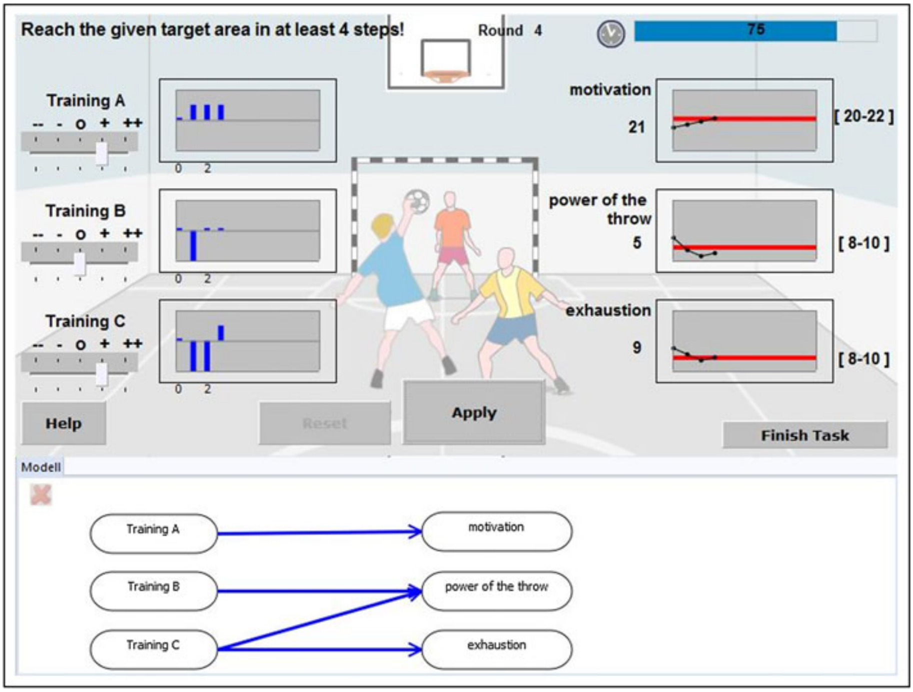
After describing the MicroDYN framework in the next section, we will introduce a versatile authoring tool and assessment platform, the CBA Item Builder, designed to fully exploit the advantages of computer-based assessment for creating innovative interactive tasks such as MicroDYN. We will illustrate the potential of computer-based simulations in an educational assessment context, for instance through easily constructing variations of simulation tasks with different properties or by scoring the application of knowledge acquisition strategies such as VOTAT. The capabilities and benefits of such a platform as a research tool will be demonstrated in an educationally motivated experimental study. This study shows how prior experience and knowledge acquired in contextualized situations influences CPS performance measured within the MicroDYN framework. We aimed to replicate the well-known effect of consistent and inconsistent of the influence of prior knowledge on CPS performance (e.g., Wittmann and Süß 1999) and to extend the basic findings with a more detailed analysis of exploration behavior.

The MicroDYN framework

As overarching approach towards measuring CPS, Greiff et al. (2012) introduced the computer-based MicroDYN tasks, which are based on the formal framework of linear structural equation (LSE) systems (Funke 2001). LSE systems have been widely applied in experimental CPS research since the 1970s (for an overview see Funke 2001). For instance, Kluge (2008) instructed participants to mix colors in her ColorSIM task and Kröner et al. (2005) developed a fictitious technical device based on an LSE system, the MultiFlux task. LSE systems contain several input and output variables, which are related to each other in a way opaque to the user. Causal relations between variables may exist either between inputs and outputs (i.e., direct effects) or between outputs (i.e., indirect effects). That is, outputs may influence each other or themselves adding a dynamic aspect independent of user's interventions to the task. Direct and indirect effects are detectable through the adequate use of exploration strategies.

Within the MicroDYN approach, a strong emphasis is put on psychometric aspects of educational assessment of CPS in contrast to the LSE systems used in experimental research over the last decades. To this end, time-on-task has been considerably shortened in MicroDYN to below 10 min per item (Greiff et al. 2012). With regard to semantic embedding, the system structure in MicroDYN can be set up in any context and labels for inputs and outputs are essentially arbitrary. Cover stories as diverse as technical machines, sports team coaching, plant growing, management of a textile factory, or feeding pets have been used. A screenshot of an example of a MicroDYN task, handball training, is depicted in Fig. 1. In this example, participants have to find out how different kinds of training (i.e., Training A, Training B, and Training C) affect the players (i.e., their motivation, their power of the throw, and their exhaustion) in order to reach a given goal state later on.

Recent empirical results on MicroDYN have contributed to establishing the construct validity of CPS. For instance, MicroDYN is significantly correlated with other CPS tasks such as the Finite State Automaton which was used in the German national extension of PISA 2000 to measure CPS, indicating convergent validity (Greiff et al. 2012). Further, MicroDYN incrementally predicted academic achievement over and above intelligence (Wüstenberg et al. 2012) and exhibited good psychometric characteristics such as strict measurement invariance (Greiff et al. in press; Wüstenberg et al. 2012). MicroDYN tasks



Reach the given target area in at least 4 steps!

Fig. 1 Screenshot of the MicroDYN-item "handball training" (knowledge application phase). The controllers of the input variables (*upper part left*) range from "− −" to "++". The current values and the target values are displayed numerically and graphically (*upper part right*). The correct causal model is presented in the *lower part*

are used in the PISA 2012 survey as means to capture CPS within an international large-scale assessment context for the first time (OECD 2010).

Even though primarily located in an educational assessment context, MicroDYN connects to cognitive-experimental research and its theories. Specifically, Novick and Bassok (2005) describe the representation of a problem (i.e., knowledge acquisition) and reaching a problem solution (i.e., knowledge application) as defining components of CPS (see above). In line with this theoretical understanding of CPS and its two overarching processes, problem solvers are instructed to perform two cognitive tasks when working on MicroDYN: to acquire knowledge and to apply this knowledge. In the first phase, knowledge acquisition, participants are instructed to explore the microworld by changing values of the inputs and to represent their gathered knowledge by drawing a causal diagram (cf. Fig. 1; Funke 2001). In the second phase, knowledge application, participants are asked to reach given target values by correctly manipulating inputs within a limited number of manipulation rounds. Furthermore, process data from both phases can be used to score strategic behavior of participants.

The MicroDYN framework allows a theoretically motivated and valid assessment of CPS yielding different performance indicators. A comprehensive description of LSE, which forms the basis of MicroDYN, can be found in Funke (2001) and the MicroDYN approach is introduced in detail in Greiff et al. (2012). We now turn to the description of a versatile technical platform for the technical implementation of MicroDYN and other interactive tasks.

A technological platform for implementing CPS tasks

Interactive CPS tasks such as MicroDYN require computer-based assessment to allow dynamic interactions between problem solver and problem. Additionally, computer-based assessment provides further benefits such as highly standardized and economical instruction and data collection, tests with adaptive difficulty levels, automatic scoring, and recording of detailed process data (cf. Scheuermann and Björnsson 2009; Wirth and Klieme 2003). Despite these benefits the adoption of computer-based testing has been lagging behind expectations (Williamson et al. 2006). While the widespread availability of computers in industrialized countries makes lack of equipment or limited computer literacy of participants less of an issue, the technical expertise required for authoring, delivering, and scoring computer-based tests remains an entrance barrier. Computer-based assessment requires the integration of different technologies and expert skills, such as programming, user interface design, server administration, and data management in addition to research expertise and content knowledge (cf. Rölke 2012). To acquire this combination of skills either requires considerable time or incurs high costs when contracted out. We therefore aimed to establish an authoring and test deployment platform that makes using computer-based assessment with interactive problem solving tasks such as MicroDYN more efficient and less time-consuming for researchers and instructors. A major requirement for such a tool was that it could be easily used by non-IT-specialists to author, test, assemble, and deliver tasks as well as provide access to response data in a form that can be further processed by common statistical software packages.

The CBA Item Builder is a generic assessment platform, which has been designed to meet these requirements. This tool is provided by the German Institute for International Educational Research (DIPF) who organizes the development of the software and collects and coordinates new requirements (for an overview see Rölke 2012). It allows users without programming experience to develop and deploy computer-based assessment tasks using a fully graphical user interface. In collaboration with the DIPF, we specified requirements to extend this generic platform for authoring and deploying interactive CPS tasks within the MicroDYN framework. The CBA Item Builder including the extensions for interactive CPS tasks are available on request and at no charge for research and non-commercial purposes (http://tba.dipf.de). The intended audience for this tool are educational researchers who wish to include CPS tasks in their assessments or instructors interested in assessing learning progress in skills such as CPS.

MicroDYN CPS tasks are created by first starting a new project in the editor of the CBA Item Builder (see Fig. 2). A number of non-interactive instruction pages containing explanatory text or images can be added to the project. The central part of each MicroDYN item is a page displaying the system that needs to be understood (knowledge acquisition) and controlled (knowledge application). In the "scenario panel" user interface elements that allow participants to interact with the system can be created and arranged using a drag-and-drop interface. The range of input and display elements available in the CBA Item Builder was specifically extended for visualizing and changing variables in MicroDYN systems, for example with line graphs and input sliders. Specific properties of these elements (e.g., limits on input ranges) can be set using property editing dialogs. The linear equation system underlying an item is graphically defined in the "modeling panel" (see Fig. 2). To measure knowledge acquisition, the runtime environment displays just the variable labels in the "modeling panel" when the task is run. Participants then draw arrows connecting the labels using the mouse to indicate their understanding of the causal relations in the system (cf. Fig. 1). In addition to interactive tasks based on the MicroDYN framework, the CBA Item Builder also provides standard response modes (e.g., rating

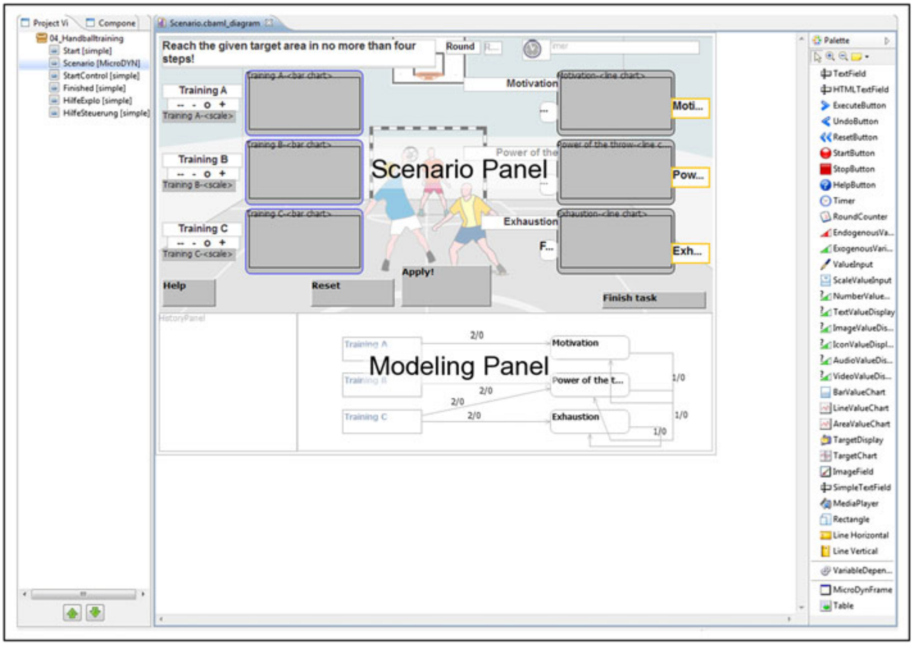

Reach the given target area in no more than four steps!

Fig. 2 Screenshot of the MicroDYN item "handball training" in the CBA Item Builder whilst under construction

scales or input boxes), multimedia capabilities to enrich stimuli (e.g., player for audio or video files), and dialog boxes to provide individual feedback to participants based on their task performance (cf. Rölke 2012).

Once the task design is completed, several tasks of different types can be combined into a test package. This test package is delivered to participants through a standard web browser. Participants either connect to a central server running the CBA delivery system or it is run locally (e.g., from a USB stick) on the computers used for testing. Tasks developed with the CBA Item Builder are also compatible with the open-source TAO platform for computer-based assessment (http://www.tao.lu).

After testing, response data can be downloaded in an XML format (http://www.w3.org/XML) from the test computer or server for further analysis. For scoring and conducting additional analysis with spreadsheets or dedicated statistics software, data can be extracted from the XML files using a transformation tool. The scoring of MicroDYN items can also be defined directly in the CBA Item Builder. A standard way of scoring in MicroDYN (e.g., Wüstenberg et al. 2012) consists of awarding credit for the correctness of the causal relations diagram drawn by participants for the knowledge acquisition score. For the knowledge application score, credit is awarded for controlling the system so that the value of an output variables falls within a target range that has been defined in the CBA Item Builder. Scoring algorithms can be easily adapted to the needs of the researcher. The CBA Item Builder allows quantitative and qualitative scoring of test performance. For instance, the deviation of target values from the values actually reached by the problem solver may serve as quantitative measure of performance in knowledge application. For a simple qualitative scoring full credit may be given if all targets are reached (e.g., for motivation, power of the throw, and exhaustion in the task "handball training", cf. Fig. 1), whereas partial credit score may indicate that only some of the targets were reached (e.g., only two out of the three variables in handball training). For a detailed description of scoring procedures please refer to the methods section of the study reported below. The resulting scores are included in the XML data files. Additionally, log files contain data about user interactions during task performance with time stamps, which may be used for further process analyses.

As an example, Table 1 presents the data contained in the XML log file for the MicroDYN item "handball training" during the exploration phase. Each round involves setting values for the input variables (Training A, B, and C) and ends with a click on the "Execute" button in the scenario. This updates the system state, including the output variables (motivation, power of throw, and exhaustion) which are displayed to the

Table 1 Example data from a log file of the MicroDYN item "handball training"

General information			Input variables			Output variables		
Time stamp	Task phase	Button pressed	Training A	Training B	Training C	Motivation	Power	Exhaustion
15:13:21	Exploration	Execute	0	0	0	15	15	15
15:13:23	Exploration	Execute	0	1	0	17	15	15
15:13:26	Exploration	Execute	0	0	1	17	17	15
15:13:29	Exploration	Execute	1	0	0	17	19	17

participant and triggers data logging in the background. The exploration pattern in this log file example illustrates a consistent application of the VOTAT strategy. The strategy allows participants to identify isolated effects of one input variable on output variables. In the example presented in Table 1, the participant changed only one input variable each round while keeping the others variables at the same value. This exploration pattern can easily be detected by automated log file analysis. The study reported below will illustrate how this type of analysis can be used to extend a more general analysis of CPS performance.

Experimental study on the relation between CPS and content knowledge

To illustrate how an easy-to-use technical platform such as the CBA Item Builder facilitates the investigation of an educationally motivated research question, we provide an experimental study as a practical example. Specifically, we used the CBA Item Builder to show the well-known effect of how domain-specific content knowledge impacts performance in CPS tasks (cf. Leutner 2002; Süß 1996). Three experimental groups worked on a CPS task that required participants to understand and manage tourist advertising to attract different kinds of tourists. The task was manipulated so that prior knowledge was either consistent with real world experiences (*consistent* group), contradictory (*inconsistent* group), or did neither help nor interfere in solving the task (*neutral* group).

In line with existing research (Funke 1992), we expected that participants in the consistent condition would outperform participants in all other conditions, because their content knowledge supports their search for information within the task environment (Klahr 2000), which is necessary to solve the problem (Kröner et al. 2005; Wüstenberg et al. 2012). Participants in the neutral group cannot rely on content knowledge and therefore have to systematically analyze the task first, yielding performance worse than the consistent group. Finally, participants in the inconsistent group were expected to perform worst, because their content knowledge and task characteristics did not match, resulting in enhanced cognitive load (Kalyuga 2009). We expected that the consistent group should significantly outperform the neutral group and the neutral group should outperform the inconsistent group (consistent > neutral > inconsistent).

Furthermore, we analyzed to what extent performance in different conditions is related to the application of a VOTAT exploration strategy and to the number of interaction rounds with the task as indicated by a detailed analysis of the log files.

Methods

Participants

The sample consisted of $N = 58$ German university students (11 male, 35 female; 12 missing sex) with a mean age of 22.11 (SD = 3.49). Students were mostly in their second year of study ($M = 1.64$, SD = .71) and received partial course credit for participation. They were randomly assigned to the consistent group ($n = 21$), the neutral group ($n = 17$), or the inconsistent group ($n = 20$).

Task

Each group worked on a MicroDYN task in which they had to explore how different kinds of tourist advertising (e.g., cultural activities) attracted different types of tourists. Tasks differed only in the semantic embedment (see Fig. 3).

In the task of the consistent group relations, input variables and output variables matched the expected content knowledge of participants based on real world experiences (e.g., increasing cultural activities led to a higher number of educational tourists; see left part of Fig. 3). In the neutral group, output variables were given generic labels without specific meaning (e.g., Tourist A, Tourist B, Tourist C; see right part of Fig. 3). In the inconsistent group, input variables and output variables used the same set of labels as the consistent group, however, labels were arbitrarily exchanged (e.g., increasing nature protection led to a higher number of party tourists; see middle part of Fig. 3). The underlying system structure, which was identical for all groups, can be modeled by the following three linear equations describing relations of input and output variables:

$$Y_{1(t+1)} = 2 \times X_{1(t)} + 2 \times X_{2(t)} + Y_{1(t)} \tag{1}$$

$$Y_{2(t+1)} = 2 \times X_{2(t)} + Y_{2(t)} \tag{2}$$

$$Y_{3(t+1)} = 2 \times X_{3(t)} + Y_{2(t)} \tag{3}$$

with t is the discrete time steps.

X_1–X_3 denote input variables and Y_1–Y_3 denote output variables. At the outset, all input variables X_1–X_3 were set on 0 and all output variables Y_1–Y_3 were set on 15. The values of the output variables changed according to the input of a participant. For instance, if a participant set X_1 on value +1 and both X_2 and X_3 on value 0 in round 1, the output variable Y_1 increased by +2 (i.e., $Y_{1(t+1)} = 2 \times 1 + 2 \times 0 + 15 = 17$; cf. Eq. 1).

The basic task with instructions, graphics, and possible interactions was created in the CBA Item Builder following the steps described in the previous section. In order to create three semantic variations that either were consistent, inconsistent, or neutral with respect to prior knowledge, we simply changed the labels for the different tourist types in the "scenario panel" (cf. Fig. 2).

Procedure and scoring

Testing sessions lasted about 20 min and started with an instruction on how to handle the user interface. Afterwards, participants were randomly assigned to three groups and

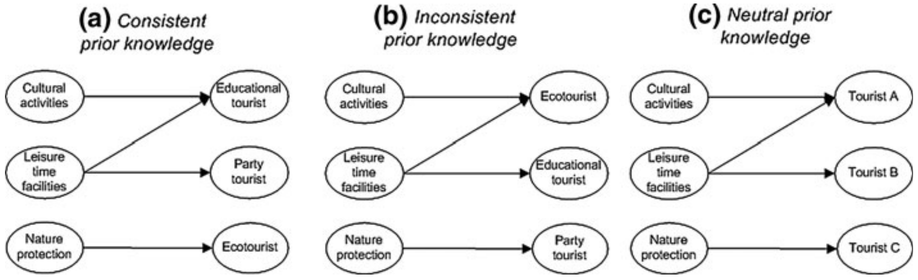


Fig. 3 Structure and semantic cover of tasks used in (**a**) consistent, (**b**) inconsistent and (**c**) neutral content knowledge group

worked on the task. In the knowledge acquisition phase, they had to explore the task and to draw a model of the causal relations of variables. In the knowledge application phase, participants had to reach given target values.

The knowledge acquisition score was calculated by subtracting the proportion of missed or falsely identified causal relations from the proportion of correctly identified causal relations, resulting in a score ranging from -1 to 1 (Frensch and Funke 1995). For instance, the correct model of the tasks used in this study consisted of four causal relations (shown in Fig. 3) out of nine possible relations between input and output variables. The knowledge acquisition score for the correct model was calculated by subtracting the proportion of missed or falsely identified causal relations (i.e., out of 5 possible) from the proportion of correctly identified causal relations (i.e., out of 4 possible). For example, when the system is correctly drawn without mistakes the resulting score is 1 ($4/4 - 0/5 = 1$). Logarithmically transformed deviations between goal state and actual state of variables were calculated separately for each of the three output variables. The average of these three logarithmic values was used as the overall knowledge application score (cf. Kluge 2008). For easier interpretation values were multiplied by -1, so that high scores indicate a smaller deviation from target values (i.e., better performance).

Results

As expected, knowledge acquisition and knowledge application scores showed on a descriptive level that the inconsistent group performed worse on both dimensions than other groups (see Table 2). However, the rank order of group means (neutral > consistent > inconsistent) was not as expected (consistent > neutral > inconsistent). We tested the statistical significance of overall group differences between consistent, neutral, and inconsistent group for both dependent variables by applying two one-way analyses of variances (ANOVA). Results revealed overall differences in performance between groups in both knowledge acquisition, $F(2, 57) = 2.86$, $p = .07$, and knowledge application, $F(2, 54) = 7.60$, $p < .001$. More specifically, linear contrasts showed that the inconsistent group performed significantly worse than the two other groups on both dimensions (knowledge acquisition: $t(55) = 2.12$, $p = .04$; knowledge application: $t(52) = 3.89$, $p < .001$), whereas no significant differences were found between consistent and neutral group (knowledge acquisition: $t(55) = 1.24$, $p = .22$; knowledge application: $t(52) = .36$, $p = .72$).

To show how process data recorded by the CBA Item Builder can be used to evaluate how content knowledge may affect performance in knowledge acquisition and knowledge

Table 2 Mean performance in knowledge acquisition and knowledge application in consistent, neutral, and inconsistent prior knowledge group

Groups	Knowledge acquisition *M* (SE)	Knowledge application *M* (SE)	*N*
Consistent	.89 (.22)	2.68 (1.41)	21
Neutral	.97 (.08)	2.87 (1.78)	17
Inconsistent	.82 (.22)	1.04 (1.55)	20

M mean score, *SE* standard error; higher values denote better performance in knowledge acquisition and knowledge application

application, we further analyzed participants' mean use of VOTAT strategy and the number of interaction rounds during exploration.

The proportion of participants using VOTAT in each experimental condition ranged from 0 (i.e., no participant used VOTAT) to 1 (i.e., all participants used VOTAT). Results showed that the inconsistent group ($M = .55$, SD $= .51$) applied significantly less often the VOTAT strategy than the neutral group ($M = 1.00$, SD $= .00$) and the consistent group ($M = .86$, SD $= .36$), indicated by a significant one-way ANOVA, $F(2, 57) = 7.30$, $p < .001$, and a significant linear contrast between inconsistent group and other groups, $t(55) = 3.67$, $p < .001$. The neutral group applied VOTAT significantly more often than the consistent group, $t(55) = 2.66$, $p = .01$.

The number of interaction rounds ranged from 0 (i.e., no interaction with the task) to 50 (i.e., maximum number of interactions with the task). The inconsistent group interacted with the task on average for more rounds (mean interaction rounds: $M_{\text{rounds}} = 9.75$, SD $= 6.16$) than the neutral group ($M_{\text{rounds}} = 7.71$, SD $= 2.23$) and the consistent group ($M_{\text{rounds}} = 7.00$, SD $= 3.07$), indicated by a marginally significant one-way ANOVA, $F(2, 57) = 2.39$, $p = .09$, and a significant linear contrast between inconsistent group and other groups, $t(55) = 2.04$, $p = .04$. No significant difference was found in the number of interactions with the task between consistent and neutral groups, $t(55) = .51$, $p = .61$.

In summary, results replicated previous research showing that participants in the inconsistent group performed worse than the other groups (Funke 1992; Klahr 2000). However, contrarily to our assumption, no significant differences were found between consistent and neutral group. Furthermore, analyses of process data revealed that the inconsistent group applied the VOTAT strategy significantly less often but interacted significantly more with the task.

Discussion

We will concentrate our discussion on how the availability of a conceptual framework and a versatile technical platform facilitated the investigation of a typical research question in CPS in the example study described above. Furthermore, we will discuss how the application of such a platform could be extended in an educational context.

In the past, researchers had only limited access to technical platforms that allowed an easy implementation, delivery, and scoring of computer-based tasks, in particular in terms of interactive simulated environments required for CPS assessment. Increasingly such platforms are becoming available to realize the first generation of computerized tests, as predicted by Bunderson et al. (1989). This is indicated by a comprehensive move towards computerized testing in combination with a decline in paper–pencil testing (Bunderson et al. 1989), as for instance in international LSAs such as PISA. The CBA Item Builder including its delivery environment is a comprehensive platform offering several advantages: (1) easy implementation of standard and innovative testing formats with a construction kit augmentable by additional modules and new task formats; (2) flexible test delivery (e.g., online and offline testing); and (3) automated scoring procedures of both pre-defined and user-defined scoring mechanisms of final and process data. This equips researchers with the means to easily share tasks, adopt existing item stems for new applications, and to directly compare different scoring procedures within a single platform.

In our experimental study we demonstrated this potential lying in a flexible platform addressing the well-known effect of prior knowledge on CPS performance. If purpose-written software had been used, the implementation of experimental conditions would have required a programmer modifying the software accordingly. Instead, contents of simple text fields were changed within the graphical text editor of the CBA Item Builder by a non-IT-specialist.

As expected, participants confronted with a CPS task contradicting their prior content knowledge performed worse than participants facing a task consistent with their knowledge or without activation of relevant content knowledge. In line with this overall pattern, the log file analysis showed that exploration of the task was more intensive in the group with inconsistent knowledge. Interestingly, the inconsistent group applied the VOTAT strategy significantly less often than the neutral and the consistent group. Apparently, mismatching prior knowledge increases exploration behavior without making it more efficient. Thus, process data indicated that prior knowledge influences how problem solvers interacted with the problem situation.

A possible explanation of these findings is that problem solving ability may be a prerequisite in acquiring content knowledge, while—once established—content knowledge attenuates the impact of problem solving strategies and strongly guides the way tasks are approached. That is, even though feedback given by the system may suggest different relations between variables, participants in the inconsistent group held on to their prior knowledge without adequately testing it, as indicated by a lower percentage of participants using VOTAT in the inconsistent group. This explanation, albeit tentative and based on a small sample, aligns well with a study of Abele et al. (2012), who showed that content knowledge was an important determinant of domain-specific problem solving at the end of a vocational training, but not at the beginning of the training when students yet relied on their generic CPS skills. This study demonstrated how the CBA Item Builder and Runtime Environment allow the use of innovative interactive task formats to investigate substantial research questions such as the influence of content knowledge on CPS performance.

Using log file data to understand the processes behind problem solving skills is a general asset of computer-based assessment. In our case, the use of the VOTAT strategy has been identified by means of the log files. Beyond exploiting process data and making the assessment more standardized and more efficient, entirely new skills such as CPS, which rely on interactively changing test environments, become accessible for research and assessment (Kyllonen 2009). From our point of view, in educational settings, in which domain-specific content knowledge has been (over-)emphasized in the past (Sternberg 1995), the focus should not exclusively lie on domain-specific problem solving skills, but comprehensively incorporate skills such as CPS (Funke 2010) as technical platforms to do so become available.

The ultimate goal of education is to foster and enhance students' personal and academic potential. According to Mayer and Wittrock (2006), this increasingly requires a shift from developing specific content knowledge towards enhancing problem solving skills relevant in several domains. In fact, much of education proceeds on the assumption of transfer (Perkins and Salomon 1989) and research by Chen and Klahr (1999) as well as Triona and Klahr (2003) suggest that students are able to use concepts taught in one context to solve new problems in other contexts. In a similar way, Fyfe et al. (2012) showed that in mathematical problem solving prior knowledge about relevant strategies but not about specific contents increases efficiency of problem exploration. In pursuit of optimizing education, we argue that assessment platforms could be used as learning environments. Specifically designed tasks may be used to assess students' domain-specific problem solving skills with specific contents (e.g., a simple physics or chemistry experiment) or on a more general level (e.g., finding a fault in a malfunctioning technical device). Results provide direct individual feedback for students and could also be used by the teacher to adapt and modify subsequent instruction. In future, the approach may be extended towards online tutoring while working on specifically designed tasks.

Moreover, individualized feedback, which plays an important role in helping students to improve their skills, is already implemented in the technical platform we presented here. Additional measures such as online tutoring may further pave the way for applications in learning contexts. Operation ARIES, an example of online tutoring in science problem solving, shows how educational measurement and individual learning can be combined into a motivating computer environment (Graesser et al. 2008). We suggest to further explore the potential of advanced technical platforms in learning environments by incorporating educational means such as feedback or online tutoring.

Accompanying learning processes in education through specifically tailored assessment using modern technical platforms is one way for educational practice to meet the changing skill demands in the 21st century. Platform availability for test implementation, delivery, and scoring in combination with a well-founded theoretical concept enables educationally motivated research, showing that now—almost four decades after the first computer-simulations were introduced in research—we may be on the threshold of a widespread application of computer-based assessment not only in the laboratory but also in classrooms across the world.

Acknowledgments This research was funded by a Grant of the German Research Foundation (DFG Fu 173/14-2), by the European Union (290683; LLLight'in'Europe), and by the German Federal Ministry of Education and Research (LSA004). We are grateful to the TBA group at DIPF (http://tba.dipf.de) for providing the authoring tool CBA Item Builder and technical support.

References

Abele, S., Greiff, S., Gschwendtner, T., Wüstenberg, S., Nickolaus, R., Nitschke, A., et al. (2012). Die Bedeutung übergreifender kognitiver Determinanten für die Bewältigung beruflicher Anforderungen. Untersuchung am Beispiel dynamischen und technischen Problemlösens [The importance of general cognitive determinants in mastering job demands. Some research on the example of dynamic and technical problem solving]. *Zeitschrift für Erziehungswissenschaft, 15*, 363–391.

Autor, D., & Dorn, D. (2009). This job is "Getting Old": Measuring changes in job opportunities using occupational age structure. *American Economic Review, 99*, 45–51.

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics, 118*, 1279–1333.

Buchner, A. (1995). Basic topics and approaches to the study of complex problem solving. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 27–63). Hillsdale, NJ: Erlbaum.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational Measurement* (pp. 367–407). New York, NY: Macmillan Publishing.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the Control of Variables Strategy. *Child Development, 70*, 1098–1120.

Danner, D., Hagemann, D., Holt, D. V., Hager, M., Schankin, A., Wüstenberg, S., et al. (2011). Measuring performance in a complex problem solving task: Reliability and validity of the Tailorshop simulation. *Journal of Individual Differences, 32*, 225–233.

Dörner, D. (1986). Diagnostik der operativen Intelligenz [Assessment of operative intelligence]. *Diagnostica, 32*(4), 290–308.

Fischer, A., Greiff, S., & Funke, J. (2012). The process of solving complex problems. *Journal of Problem Solving, 4*(1), 19–42.

Frensch, P. A., & Funke, J. (Eds.). (1995). *Complex problem solving: The European perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Funke, J. (1992). Dealing with dynamic systems: Research strategy, diagnostic approach and experimental results. *German Journal of Psychology, 16*, 24–43.

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking and Reasoning, 7*(1), 69–89.

Funke, J. (2003). *Problemlösendes Denken [Problem solving]*. Stuttgart: Kohlhammer.

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

Fyfe, E. R., Rittle-Johnson, B., & DeCaro, M. S. (2012). The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters. *Journal of Educational Psychology, 104*, 1094–1108.

Gonzalez, C., Thomas, R. P., & Vanyukov, P. (2005). The relationships between cognitive ability and dynamic decision making. *Intelligence, 33*(2), 169–186.

Graesser, A. C., Jeon, M., & Dufty, D. (2008). Agent technologies designed to facilitate interactive knowledge construction. *Discourse Processes, 45*, 298–322.

Greiff, S., & Fischer, A. (2013). Der Nutzen einer Komplexen Problemlösekompetenz: Theoretische Überlegungen und empirische Befunde [Usefulness of Complex Problem Solving competency: Theoretical considerations and empirical results]. *Zeitschrift für Pädagogische Psychologie, 27*(1), 1–13.

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic Problem Solving: A new measurement perspective. *Applied Psychological Measurement, 36*, 189–213.

Greiff, S., Wüstenberg, S., Molnar, G., Fischer, A., Funke, J., & Csapo, B. (in press). Complex Problem Solving in educational settings—something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology*. doi:10.1037/a0031856.

Holyoak, K. J. (1985). The pragmatics of analogical transfer. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 59–87). New York, NJ: Academic Press.

Kalyuga, S. (2009). Knowledge elaboration: A cognitive load perspective. *Learning and Instruction, 19*, 402–410.

Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*. Cambridge, MA: MIT Press.

Kluge, A. (2008). Performance assessment with microworlds and their difficulty. *Applied Psychological Measurement, 32*(2), 156–180.

Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence, 33*, 347–368.

Kyllonen, P. C. (2009). New constructs, methods, and directions for computer-based assessment. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment* (pp. 151–156). Luxembourg: Office for Official Publications of the European Communities.

Leutner, D. (2002). The fuzzy relationship of intelligence and problem solving in computer simulations. *Computers in Human Behavior, 18*, 685–697.

Mayer, R. E. (2003). *Learning and instruction*. Upper Saddle River, NJ: Prentice Hall.

Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 287–303). Mahwah, NJ: Lawrence Erlbaum.

Novick, L. R., & Bassok, M. (2005). Problem solving. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 321–349). Cambridge, NY: University Press.

OECD (2010). *PISA 2012 field trial problem solving framework*. Accessed at: http://www.oecd.org/dataoecd/8/42/46962005.pdf.

OECD (2012). *Better skills, better jobs, better lives. Highlights of the OECD skills strategy*. Accessed at: http://www.oecd-ilibrary.org/docserver/download/fulltext/9112165e5.pdf.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher, 18*(10), 16–25.

Raven, J. (2000). Psychometrics, cognitive ability, and occupational performance. *Review of Psychology, 7*, 51–74.

Rölke, H. (2012). The item builder: A graphical authoring system for complex item development. in *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education* (pp. 344–353). Chesapeake, VA: AACE.

Scheuermann, F., & Björnsson, J. (2009). *The transition to computer-based assessment*. Luxembourg: Office for Official Publications of the European Communities.

Smith, M. C., & Reio, T. G. (2006). Adult development, schooling, and the transition to work. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 115–138). Mahwah, NJ: Lawrence Erlbaum.

Spitz-Oener, A. (2006). Technical change, job tasks, and rising educational demands: Looking outside the wage structure. *Journal of Labor Economics, 24*(2), 235–270.

Sternberg, R. J. (1995). Expertise in complex problem solving: A comparison of alternative concepts. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving. The European perspective* (pp. 295–321). Hillsdale, NJ: Lawrence Erlbaum.

Süß, H.-M. (1996). *Intelligenz, Wissen und Problemlösen*. Göttingen: Hogrefe.

Triona, L. M., & Klahr, D. (2003). Point and click or grab and heft: Comparing the influence of physical and virtual instructional materials on elementary school students' ability to design experiments. *Cognition and Instruction, 21*, 149–173.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development, 51*, 1–10.

Vollmeyer, R., Burns, B. D., & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science, 20*, 75–100.

Williamson, D. M., Bejar, I. I., & Mislevy, R. J. (2006). *Automated scoring of complex tasks in computer-based testing: An introduction*. Mahwah, NJ: Lawrence Erlbaum.

Wirth, J., & Klieme, E. (2003). Computer-based assessment of problem solving competence. *Assessment in Education: Principles, Policy & Practice, 10*(3), 329–345.

Wittmann, W. W., & Süß, H. M. (1999). Investigating the paths between working memory, intelligence, knowledge and complex problem solving: Performances via Brunswik-symmetry. In P. L. Ackermann, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait and content* (pp. 77–108). Washington, DC: American Psychological Association.

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving. More than reasoning? *Intelligence, 40*, 1–14.

Samuel Greiff is senior researcher and ATTRACT fellow at the University of Luxembourg. His main interests lie in educational measurement of complex mental abilities.

Sascha Wüstenberg is research associate at the University of Luxembourg. His main interest lies in educational measurement of problem solving competency.

Daniel V. Holt is research associate at the University of Heidelberg specializing in problem solving and executive functions.

Frank Goldhammer is full professor for educational and psychological assessment with focus on technology-based assessment at the German Institute for International Educational Research (DIPF) and the Centre for International Student Assessment (ZIB).

Joachim Funke is full professor of general and theoretical psychology at the University of Heidelberg with a research focus on cognitive processing and problem solving.