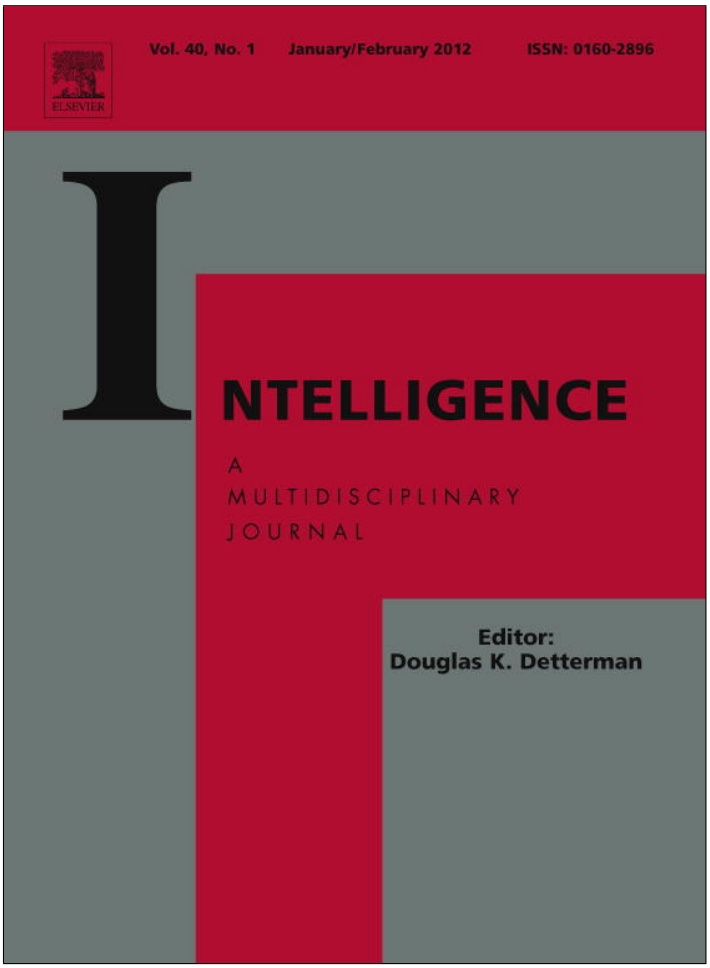

Intelligence 40 (2012) 1–14

Contents lists available at SciVerse ScienceDirect

Intelligence

Complex problem solving — More than reasoning?

Sascha Wüstenberg, Samuel Greiff*, Joachim Funke

Department of Psychology, University of Heidelberg, Germany

ARTICLE INFO

Article history:
Received 12 July 2011
Received in revised form 2 November 2011
Accepted 10 November 2011
Available online 3 December 2011

Keywords:
Complex problem solving
Intelligence
Dynamic problem solving
MicroDYN
Linear structural equations
Measurement

ABSTRACT

This study investigates the internal structure and construct validity of Complex Problem Solving (CPS), which is measured by a *Multiple-Item-Approach*. It is tested, if (a) three facets of CPS – *rule identification* (adequateness of strategies), *rule knowledge* (generated knowledge) and *rule application* (ability to control a system) – can be empirically distinguished, how (b) reasoning is related to these CPS-facets and if (c) CPS shows incremental validity in predicting school grade point average (GPA) beyond reasoning. N = 222 university students completed MicroDYN, a computer-based CPS test and Ravens Advanced Progressive Matrices. Analysis including structural equation models showed that a 2-dimensionsal model of CPS including *rule knowledge* and *rule application* fitted the data best. Furthermore, reasoning predicted performance in *rule application* only indirectly through its influence on *rule knowledge* indicating that learning during system exploration is a prerequisite for controlling a system successfully. Finally, CPS explained variance in GPA even beyond reasoning, showing incremental validity of CPS. Thus, CPS measures important aspects of academic performance not assessed by reasoning and should be considered when predicting real life criteria such as GPA.

General intelligence is one of the most prevalent constructs among psychologists as well as non-psychologists (Sternberg, Conway, Ketron, & Bernstein, 1981) and frequently used as predictor of cognitive performance in many different domains, e.g., in predicting school success (Jensen, 1998a), life satisfaction (Eysenck, 2000; Sternberg, Grigorenko, & Bundy, 2001) or job performance (Schmidt & Hunter, 2004). However, considerable amount of variance in these criteria remains unexplained by general intelligence (Neisser et al., 1996). Therefore, Rigas, Carling, and Brehmer (2002) suggested the use of microworlds (i.e., computer-based complex problem solving scenarios) to increase the predictability of job related success. Within complex problem solving (CPS) tasks, people actively interact with an unknown system consisting of many highly interrelated variables and are asked to actively generate knowledge to achieve certain goals (e.g., managing a *Tailorshop*; Funke, 2001). In this paper, we argue that previously used measurement devices of CPS suffer from a methodological point of view. Using a newly developed approach, we investigate (1) the internal structure of CPS, (2) how CPS is related to reasoning — which is seen as an excellent marker of general intelligence (Jensen, 1998b) — and (3) if CPS shows incremental validity even beyond reasoning.

1. Introduction

Reasoning can be broadly defined as the process of drawing conclusions in order to achieve goals, thus informing problem-solving and decision-making behavior (Leighton, 2004). For instance, reasoning tasks like the *Culture Fair Test* (CFT-20-R; Weiß, 2006) or Ravens *Advanced Progressive Matrices* (APM; Raven, 1958) require participants to identify and acquire rules, apply them and coordinate two or more rules in order to complete a problem based on visual patterns (Babcock, 2002). Test performance on APM has been suggested to be dependent on executive control processes that allow a subject to analyze complex problems, assemble solution strategies, monitor performance and adapt behavior as

* Corresponding author at: Department of Psychology, University of Heidelberg, Hauptstraße 47–51, 69117 Heidelberg, Germany. Tel.: +49 6221 547613; fax: +49 547273.
E-mail address: Samuel.greiff@psychologie.uni-heidelberg.de (S. Greiff).

0160-2896/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2011.11.003

testing proceeds (Marshalek, Lohman, & Snow, 1983; Wiley, Jarosz, Cushen, & Colflesh, 2011).

However, the skills linked to executive control processes within reasoning and CPS are often tagged with the same labels: Also in CPS, acquiring and applying knowledge and monitoring behavior are seen as important skills in order to solve a problem (Funke, 2001), e.g., while dealing with a new type of mobile phone. For instance, if a person wants to send a text message for the first time, he or she will press buttons in order to navigate through menus and get feedback. Based on the feedback he or she persists in or changes behavior according to how successful the previous actions have been. This type of mobile phone can be seen as a CPS-task: The problem solver does not know how several variables in a given system (e.g., mobile phone) are connected with each other. His or her task is to gather information (e.g., by pressing buttons to toggle between menus) and to generate knowledge about the system's structure (e.g., the functionality of certain buttons) in order to reach a given goal state (e.g., sending a text message). Thus, elaborating and using appropriate strategies in order to solve a problem is needed in CPS and as well in reasoning tasks like APM (Babcock, 2002), so that Wiley et al. (2011) name APM a visuospatial reasoning and problem solving task.

However, are the underlying processes while solving static tasks like APM really identical to complex and interactive problems, like in the mobile phone example? And does reasoning assess performance in dealing with such problems? Raven (2000) denies that and points towards different demands upon the problem solver while dealing with problem solving tasks as compared to reasoning tasks.

> *…It* [Problem solving] *involves initiating, usually on the basis of hunches or feelings, experimental interactions with the environment to clarify the nature of a problem and potential solutions. […] In this way they [the problem solvers] can learn more about the nature of the problem and the effectiveness of their strategies. […] They can then modify their behaviour and launch a further round of experimental interactions with the environment (Raven, 2000, p. 479).*

Raven (2000) separates CPS from reasoning assessed by APM. He focuses on dynamic interactions necessary in CPS for revealing and incorporating previously unknown information as well as achieving a goal using subsequent steps which depend upon each other. This is in line with Buchner's understanding (1995) of complex problem solving (CPS) tasks:

> *Complex problem solving (CPS) is 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 (Buchner, 1995, p. 14).*

The main differences between reasoning tasks and CPS tasks are that in the latter case (1) not all information necessary to solve the problem is given at the outset, (2) the problem solver is required to actively generate information via applying adequate strategies, and (3) procedural abilities have to be used in order to control a given system, such as when using feedback in order to persist or change behavior or to counteract unwanted developments initiated by the system (Funke, 2001). Based on these different demands upon the problem solver, Funke (2010) emphasized that CPS requires not only a sequence of simple cognitive operations, but complex cognition, i.e., a series of different cognitive operations like action planning, strategic development, knowledge acquisition and evaluation, which all have to be coordinated to reach a certain goal.

In summary, on a conceptual level, reasoning and CPS both assess cognitive abilities necessary to generate and apply rules, which should yield in correlations between both constructs. Nevertheless, according to the different task characteristics and cognitive processes outlined above, CPS should also show divergent validity to reasoning.

1.1. Psychometrical considerations for measuring CPS

Numerous attempts have been made to discover the relationship between CPS and reasoning empirically (for an overview see, e.g., Beckmann, 1994; Beckmann & Guthke, 1995; Funke, 1992; Süß, 1996; Wirth, Leutner, & Klieme, 2005). Earlier CPS-research in particular reported zero-correlations (e.g., Joslyn & Hunt, 1998; Putz-Osterloh, 1981), while more recent studies revealed moderate to high correlations between CPS and reasoning (e.g., Wittmann & Hattrup, 2004; Wittmann & Süß, 1999). For instance, Gonzalez, Thomas, and Vanyukov (2005) showed that performance in the CPS-scenarios *Water Purification Plant* (0.333, $p<0.05$) and *Firechief* (0.605; $p<0.05$) were moderately to highly correlated with APM.

In order to explain the incongruity observed, Kröner, Plass, and Leutner (2005) summarized criticisms of various authors on CPS research (e.g., Funke, 1992; Süß, 1996) and stated, that the relationship between CPS and reasoning scenarios could only be evaluated meaningfully if three general conditions were fulfilled.

1.1.1. Condition (A): Compliance with requirements of test theory

Early CPS work (Putz-Osterloh, 1981) suffered particularly from a lack of reliable CPS-indicators, leading to low correlations of CPS and reasoning (Funke, 1992; Süß, 1996). If reliable indicators were used, correlations between reasoning and CPS increased significantly (Süß, Kersting, & Oberauer, 1993) and CPS even predicted supervisor ratings (Danner et al., 2011). Nevertheless, all studies mentioned above used scenarios in which problem solving performance may be confounded with prior knowledge leading to condition (B).

1.1.2. Condition (B): No influence of simulation-specific knowledge acquired under uncontrolled conditions

Prior knowledge may inhibit genuine problem solving processes and, hence, negatively affect the validity of CPS. For instance, this applies to the studies of Wittmann and Süß (1999), who claimed CPS to be a conglomerate of knowledge and intelligence. In their study, they assessed reasoning (subscale processing capacity of the *Berlin Intelligence Structure Test — BIS-K*; Jäger, Süß, & Beauducel, 1997) and measured CPS by three different tasks (*Tailorshop*, *PowerPlant*, *Learn*). Performance between these CPS tasks was correlated.

However, correlations vanished when system-specific knowledge and reasoning were partialled out. The authors' conclusion of CPS being only a conglomerate is questionable, because the more prior knowledge is helpful in a CPS task, the more this knowledge will suppress genuine problem solving processes like searching for relevant information, integrating knowledge or controlling a system (Funke, 2001). In order to avoid these uncontrolled effects, CPS scenarios which do not rely on domain-specific knowledge ought to be used.

1.1.3. Condition (C): Need for an evaluation-free exploration phase

An exploration phase for identifying the causal connections between variables should not contain any target values to be reached in order to allow participants to have an equal opportunity to use their knowledge-acquisition abilities under standardized conditions (Kröner et al., 2005).

Consequently, Kröner et al. (2005) designed a CPS scenario based on linear structural equation systems (Funke, 2001) called *MultiFlux* and incorporated the three suggestions outlined. Within *MultiFlux*, participants first explore the task and their generated knowledge is assessed. Participants then are presented the correct model of the causal structure and asked to reach given target values. Finally, three different facets of CPS are assessed — the use of adequate strategies (*rule identification*), the knowledge generated (*rule knowledge*) and the ability to control the system (*rule application*). Results showed that reasoning (measured by BIS-K) predicted each facet (*rule identification*: $r = 0.48$; *rule knowledge* $r = 0.55$; *rule application* $r = 0.48$) and the prediction of *rule application* by reasoning was even stronger than the prediction of *rule application* by *rule knowledge* ($r = 0.37$). In a more recent study using *MultiFlux*, Bühner, Kröner, and Ziegler (2008) extended the findings of Kröner et al. (2005). They showed that in a model containing working memory (measured by a spatial coordination task; Oberauer, Schulze, Wilhelm, & Süß, 2005), CPS and intelligence (measured by *Intelligence Structure Test 2000 R*; Amthauer, Brocke, Liepmann, & Beauducel, 2001), intelligence predicted each CPS-facet (*rule knowledge* $r = 0.26$; *rule application* $r = 0.24$; *rule identification* was not assessed), while the prediction of *rule application* by *rule knowledge* was not significant ($p > 0.05$). In both studies, reasoning predicted rule application more strongly than rule knowledge did. Thus, the authors concluded that *MultiFlux* can be used as a measurement device for the assessment of intelligence, because each facet of CPS can be directly predicted by intelligence (Kröner et al., 2005).

In summary, Kröner et al. (2005) pointed towards the necessity of measuring CPS in a test-theoretical sound way and developed a promising approach based on three conditions. Nevertheless, some additional methodological issues that may influence the relationship between reasoning and CPS were not sufficiently regarded.

1.2. Prerequisite — Multiple-item-testing

MultiFlux, as well as all other CPS scenarios previously mentioned may be considered *One-Item-Tests* (Greiff, in press). These scenarios generally consist of one specific system configuration (i.e., variables as well as relations between them remain the same during test execution). Thus, all indicators assessing *rule knowledge* gained during system exploration are related to the very same system structure and consequently depend on each other. This also accounts for indicators of *rule application*: Although participants work on a series of independent *rule application* tasks with different target goals, these tasks also depend on the very same underlying system structure. Consequently, basic test theoretical assumptions are violated making CPS scenarios comparable to an intelligence test with one single item, but with multiple questions on it. The dimensionality of the CPS construct cannot be properly tested, because indicators within each of the dimensions *rule knowledge* and *rule application* are dependent on each other. Thus, *One-Item-Testing* inhibits a sound testing of the dimensionality of CPS.

There are two different ways to assess *rule application* in CPS tasks, either by implementing (a) only one control round or (b) multiple control rounds. Using (a) only one control round enhances the influence of reasoning on *rule application*. For instance, within *MultiFlux* (Bühner et al., 2008; Kröner et al., 2005), *rule application* is assessed by participants' ability to properly set controls in all input variables in order to achieve given target values of output variables within one control round. During these tasks, no feedback is given to participants. Thus, procedural aspects of *rule application* like using feedback in order to adjust behavior or counteract system changes not directly controllable by the problem solver are not assessed. Because of this lack of interaction between problem solver and problem, *rule application* in *MultiFlux* assesses primarily cognitive efforts in applying rules also partly measured in reasoning tasks — and less procedural aspects genuine to CPS. Additionally, within *MultiFlux*, *rule knowledge* tasks are also similar to *rule application* tasks, because knowledge is assessed by predicting values of a subsequent round given that input variables were in a specific configuration at the round before. This kind of knowledge assessment requires not only knowledge about rules, but also the ability to apply rules in order to make a prediction. Consequently, *rule knowledge* and *rule application* as well as reasoning and *rule application* were strongly correlated ($r = 0.77$ and $r = 0.51$, respectively; Kröner et al., 2005). However, if intelligence was added as a predictor of both *rule knowledge* and *rule application*, the path between *rule knowledge* and *rule application* was significantly lowered ($r = 0.37$; Kröner et al., 2005) or even insignificant (Bühner et al., 2008). This shows that *rule application* assessed by one-step control rounds measures similar aspects of CPS as rule knowledge — and these aspects depend on reasoning to a comparable extent, reducing the validity of the construct CPS. Thus, multiple control rounds have to be used in order to also allow the assessment of CPS abilities like using and incorporating feedback in *rule application*.

However, using (b) multiple control rounds does not solve the problem within *One-Item-Testing*, because that would lead to confounded indicators of *rule application*: As long as *rule application* tasks are based on the same system structure, participants may use given feedback and gather additional knowledge (improved *rule knowledge*) during subsequently administered *rule application* tasks. Consequently, within *rule application*, not only the ability to control a system would be measured, but also the ability to gain further knowledge about its structure (Bühner et al., 2008).

Thus, the only way to assess CPS properly, enabling direct interaction and inhibiting confounded variables, is by adding a prerequisite (D) – the use of multiple items differing in system configuration – to the three conditions (A–C) Kröner et al. (2005) mentioned for a proper assessment of CPS. In a *Multiple-Item-Approach*, multiple (but limited) control rounds can be used, because additional knowledge that is eventually gained during *rule application* does not support participants in the following item based on a completely different structure.

Besides using a *Multiple-Item-Approach*, we also want to include external criteria of cognitive performance (e.g., school grade) in order to check construct validity of CPS. Research that has done so far mostly tested exclusively the predictive validity of system control, i.e. *rule application* (e.g., Gonzalez, Vanyukov, & Martin, 2005). This is surprising, because according to Buchner's (1995) definition as well as Raven's (2000), the aspects of actively using information (*rule identification*) in order to generate knowledge (*rule knowledge*) also determine the difference between reasoning and CPS — and not only the application of rules. Consequently, predictive and incremental validity of all relevant CPS facets should be investigated.

In summary, the aim of this study is to re-evaluate as well as to extend some questions raised by Kröner et al. (2005):

(1) Can the three facets of CPS still be empirically separated within a *Multiple-Item-Approach*? Thus, the dimensionality of the construct CPS will be under study, including a comparison between a multi- and a unidimensional (and more parsimonious) model, which has not been done yet.
(2) Is CPS only another measure of reasoning? This question includes the analysis of which CPS facets can be predicted by reasoning and how they are related.
(3) Can CPS be validated by external criteria? This question targets the predictive and incremental validity of each CPS facet.

1.3. The MicroDYN-approach

The MicroDYN-approach, aimed at capturing CPS, incorporates the prerequisites mentioned above (see Greiff, in press). In contrast to other CPS scenarios, MicroDYN uses multiple and independent items to assess CPS ability. A complete test set contains 8 to 10 minimal but sufficiently complex items, each lasting about 5 min, in their sum a total testing time of less than 1 h including instruction. MicroDYN-items consist of up to 3 input variables (denoted by A, B and C), which can be related to up to 3 output variables (denoted by X, Y and Z; see Fig. 1).

Input variables influence output variables, where only the former can be actively manipulated by the problem solver. There are two kinds of connections between variables: Input variables which influence output variables and output variables which influence themselves. The latter may occur if different output variables are related (side effect; see Fig. 1: Y to Z) or if an output variable influences itself (autoregressive process; see Fig. 1: X to X).

MicroDYN-tasks can be fully described by linear structural equations (for an overview see Funke, 2001), which have been used in CPS research to describe complex systems since the early 1980ies. The number of equations necessary to describe all possible relations is equal to the number of output variables. For the specific example in Fig. 1, Eqs. (1) to (3) are needed:

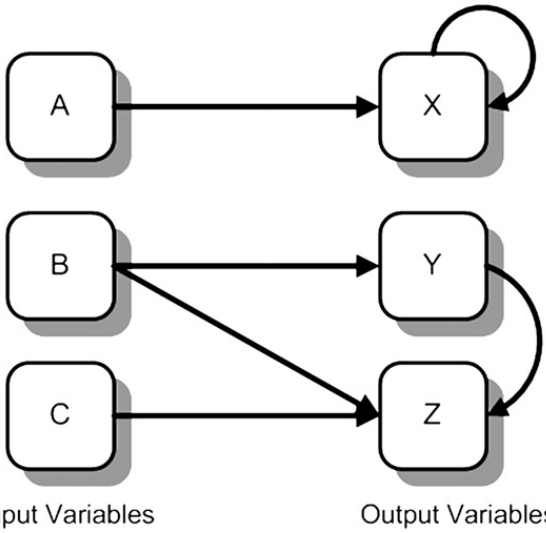

Fig. 1. Structure of a typical MicroDYN item displaying 3 input (A, B, C) and 3 output (X, Y, Z) variables.

$$X_{(t+1)} = a_1 * A_{(t)} + a_2 * X_{(t)} \quad (1)$$

$$Y_{(t+1)} = a_3 * B_{(t)} + Y_{(t)} \quad (2)$$

$$Z_{(t+1)} = a_4 * B_{(t)} + a_5 * C_{(t)} + a_6 * Y_{(t)} + Z_{(t)} \quad (3)$$

with t = discrete time steps, a_i = path coefficients, $a_i \neq 0$, and $a_2 \neq 1$.

Within each MicroDYN-item, the path coefficients are fixed to a certain value (e.g., $a_1 = +1$) and participants may vary variable A, B and C. Although Fig. 1 may look like a path diagram and the linear equations shown above may look like a regression model, both illustrations only show how inputs and outputs are connected within a given system.

Different cover stories were implemented for each item in MicroDYN (e.g. feeding a cat, planting pumpkins or driving a moped). In order to avoid uncontrolled influences of prior knowledge, variables were either labeled without deep semantic meaning (e.g., *button A*) or fictitiously (e.g., *sungrass* as name for a flower). For instance, in the item "handball" (see Fig. 2; for linear structural equations see Appendix A), different kinds of training labeled training A, B and C served as input variables whereas different team characteristics labeled motivation, power of throw, and exhaustion served as output variables.

While working on MicroDYN, participants face three different tasks that are directly related to the three facets of problem solving ability considered by Kröner et al. (2005). In the exploration phase, (1) participants freely explore the system and are asked to discover the relationships between the variables involved. Here, the adequateness of their strategies is assessed (facet *rule identification*). For instance, in the handball training item, participants may vary solely the value of training A in round 1 by manipulating a slider (e.g., from "0" to "++"). After clicking on the "apply"-button,

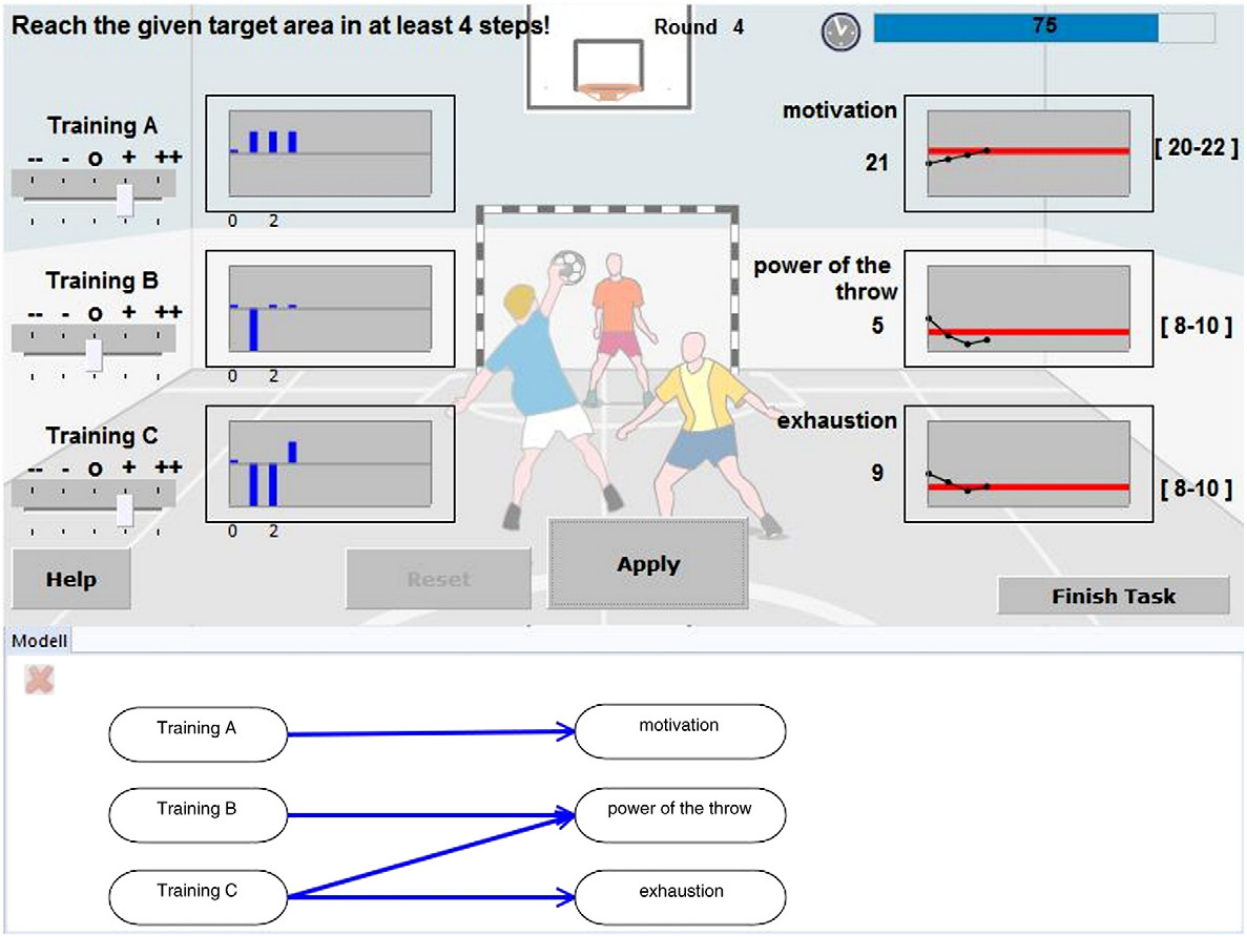

Fig. 2. Screenshot of the MicroDYN-item "handball training" control phase. The controllers of the input variables range from "- -" (value = −2) to "++" (value = +2). The current value is displayed numerically and the target values of the output variables are displayed graphically and numerically.

they will see how the output variables change (e.g., value on motivation increases).

Simultaneously, (2) participants have to draw lines between variables in a causal model as they suppose them to be, indicating the amount of generated knowledge (facet *rule knowledge*). For instance, participants may draw a line between training A and motivation by merely clicking on both variable names (see model at the bottom of Fig. 2). Afterwards, in the control phase, (3) participants are asked to reach given target goals in the output variables within 4 steps (facet *rule application*). For instance, participants have to increase the value of motivation and power of the throw, but minimize exhaustion (not displayed in Fig. 2). In order to disentangle *rule knowledge* and *rule application*, the correct model is given to the participants during *rule application*. Within each item, the exploration phase assessing *rule identification* and *rule knowledge* lasts about 180 s and the control phase lasts about 120 s.

1.4. The present study

1.4.1. Research question (1): Dimensionality

Kröner et al. (2005) showed that three different facets of CPS ability, *rule identification*, *rule knowledge* and *rule application* can be empirically distinguished. However, all indicators derived are based on one single item, leading to dependencies of indicators incompatible with psychometrical standards. Thus, the dimensionality of CPS has to be tested in a *Multiple-Item-Approach* with independent performance indicators.

Hypothesis (1). The indicators of *rule identification*, *rule knowledge* and *rule application* load on three corresponding factors. A good fit of the 3-dimensional model in confirmatory factor analysis (CFA) is expected. Comparisons with less dimensional (and more parsimonious) models confirm that these models fit significantly worse.

1.4.2. Research question (2): CPS and reasoning

According to the theoretical considerations raised in the Introduction, reasoning and CPS facets should be empirically related. In order to gain more specific insights about this connection, we assume that the process oriented model shown in Fig. 3 is appropriate to describe the relationship between reasoning and different facets of CPS.

In line with Kröner et al. (2005), we expect *rule identification* to predict *rule knowledge* (path a), since adequate use of strategies yields better knowledge of causal relations. *Rule knowledge* predicts *rule application* (path b), since knowledge about causal relations leads to better performance in controlling a system. Furthermore, reasoning should predict performance in *rule identification* (path c) and *rule knowledge* (path d), because more intelligent persons are expected to better explore any given system and to acquire more system knowledge. However, we disagree with Kröner et al. (2005) in our

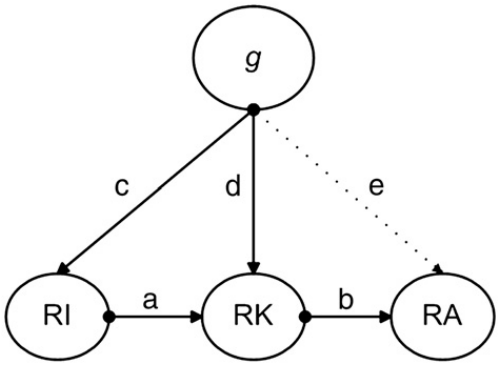

Fig. 3. Theoretical model of the relations between reasoning (g) and the CPS facets rule identification (RI), rule knowledge (RK) and rule application (RA). The dotted line indicates a insignificant path coefficient (e). All four other paths are expected to be significant.

predictions that reasoning directly predicts performance in *rule application*. In their results, the direct path (e) indicated that irrespectively of the amount of *rule knowledge* acquired beforehand, more intelligent persons used the correct model given in the control phase to outperform less intelligent ones in *rule application*. We assume that this result is due to the way *rule application* was assessed in *MultiFlux*. Participants had to reach certain target values as output variables within one single round. Thus, procedural abilities (e.g., using feedback in order to adjust behavior during system control) were not necessary and *rule application* solely captured abilities also assessed by reasoning. This leads to a significant path (e) and reduced the impact of path (b) (Bühner et al., 2008; Kröner et al., 2005). As outlined above, using multiple control rounds within a *One-Item-Approach* leads to confounded variables of rule knowledge and rule application. A *Multiple-Item-Approach*, however, allows multiple independent control rounds forcing participants to use procedural abilities (not assessed by reasoning) in order to control the system.

Consequently, learning to handle the system during exploration is essential and analysis of the correct model given in the control phase is not sufficient for system control. Thus, more intelligent participants should only be able to outperform less intelligent ones in *rule application*, because they have gained more system knowledge and have better procedural abilities necessary for *rule application*. Reasoning should predict performance in *rule application*, however, only indirectly via its influence on *rule identification* and *rule knowledge* (indicated by an insignificant direct effect in path e).

Hypothesis (2). The theoretical process model (shown in Fig. 3) is empirically supported, indicating that *rule identification and rule knowledge* fully mediate the relationship between reasoning and *rule application*.

1.4.3. Research question (3): Predictive and incremental validity of CPS

Finally, we assume that CPS facets predict performance in important external criteria like *school grade point average* (GPA) even beyond reasoning indicating the incremental validity of CPS. The ability to identify causal relations and to gain knowledge when confronted with unknown systems is frequently demanded in different school subjects (OECD, 2004). For instance, tasks in physics require analyzing elementary particles and their interactions in order to understand the properties of a specific matter or element. However, actively controlling a system by using procedural abilities is less conventional at school. Consequently, a significant prediction of GPA by *rule identification* and *rule knowledge* is expected, whereas *rule application* should be a less important predictor.

Hypothesis (3). CPS ability measured by the CPS facets *rule identification* and *rule knowledge* significantly predict GPA beyond reasoning, whereas there is no increment in prediction for *rule application*.

2. Method

2.1. Participants

Participants were 222 undergraduate and graduate students (154 female, 66 male, 2 missing sex; age: M = 22.8; SD = 4.0), mainly from social sciences (69%, thereof 43% studying psychology) followed by natural sciences (14%) and other disciplines (17%). Most of the students were undergraduates (n = 208). Students received partial course credit for participation and an additional 5 € (approx. 3.5 US $) if they worked conscientious. A problem solver was treated as working not conscientiously, if more than 50% data were missing on APM and if the mean of the exploration rounds in MicroDYN was less than three rounds. Within MicroDYN, at least three rounds are needed to identify all causal relations in an item. We excluded participants from the analyses either because they were not working conscientiously (n = 4) or because of missing data occurring due to software problems (e.g., data was not saved properly; n = 12). Finally, data for 222 students were available for the analyses. The study took place at the Department of Psychology at the University of Heidelberg, Germany.

2.2. Materials

2.2.1. MicroDYN

Testing of CPS was entirely computer-based. Firstly, participants were provided with a detailed instruction including two items in which they actively explored the surface of the program and were informed about what they were expected to do: gain information about the system structure (*rule identification*), draw a model (*rule knowledge*) and finally control the system (*rule application*). Subsequently, participants dealt with 8 MicroDYN items. The task characteristics (e.g., number of effects) were varied in order to produce items across a broad range of difficulty (Greiff & Funke, 2010; see section on MicroDYN approach and also Appendix A for equations).

2.2.2. Reasoning

Additionally, participants' reasoning ability was assessed using a computer adapted version of the *Advanced Progressive Matrices* (APM, Raven, 1958). This test has been extensively standardized for a population of university students and is seen as a valid indicator of fluid intelligence (Raven, Raven, & Court, 1998).

2.2.3. GPA

Participants provided demographical data and their GPA in self-reports.

3. Design

Test execution was divided into two sessions, each lasting approximately 50 min. In session 1, participants worked on MicroDYN. In session 2, APM was administered first and participants provided demographical data afterwards. Time between sessions varied between 1 and 7 days (M = 4.2, SD = 3.2).

3.1. Dependent variables

In MicroDYN, ordinal indicators were used for each facet. This is in line with Kröner et al. (2005), but not with other research on CPS that uses indicators strongly depending on single system characteristics (Goode & Beckmann, 2011; Klieme, Funke, Leutner, Reimann, & Wirth, 2001). However, ordinal indicators can be used to measure interval-scaled latent variables within structural equation modeling approach (SEM; Bollen, 1989) and also allow analyses of all items within item response theory (IRT; Embretson & Reise, 2000).

For *rule identification,* full credit was given if participants showed a consistent use of VOTAT (i.e., vary one thing at a time; Vollmeyer & Rheinberg, 1999) for all variables. The use of VOTAT enables the participants to identify the isolated effect of one input variable on the output variables (Fig. 1). Participants were assumed to have mastered VOTAT when they applied it to each input variable at least once during exploration. VOTAT is seen as the best strategy to identify causal relations within linear structural equation systems (Tschirgi, 1980) and frequently used in CPS research as indicator of an adequate application of strategies (e.g., Burns & Vollmeyer, 2002; Vollmeyer, Burns, & Holyoak, 1996). Another possible operationalization of rule identification is to assess self-regulation abilities of problem solvers as introduced by Wirth (2004) and Wirth and Leutner (2008) using the scenario *Space Shuttle*. Their indicator is based on the relation of generating and integrating information while exploring the system. Generating information means to perform an action for the first time, whereas integrating information means to perform the same actions that had previously been done once again to check whether the relationships of input and output variables had been understood correctly. An appropriate self-regulation process is indicated by focussing on generating new information in the first rounds of an exploration phase and by focussing on integrating of information in the latter rounds. However, this kind of operationalization is more efficient in tasks, in which working memory limits the ability to keep all necessary information in mind. Within MicroDYN, participants are allowed to simultaneously track the generated information by drawing a model, rendering the process of integrating information less essential. Thus, we only used VOTAT as an indicator of rule identification.

For *rule knowledge*, full credit was given if the model drawn was completely correct and in case of *rule application*, if target areas of all variables were reached. A more detailed scoring did not yield any better results on psychometrics. Regarding APM, correct answers in Set II were scored dichotomously, accordingly to the recommendation in the manual (Raven et al., 1998).

3.2. Statistical analysis

To analyze data we ran CFA within the structural equation modeling approach (SEM; Bollen, 1989) and Rasch analysis within item response theory (IRT). We used the software MPlus 5.0 (Muthén & Muthén, 2007a) for SEM calculations and Conquest 3.1 for Rasch analysis (Wu, Adams, & Haldane, 2005). Descriptive statistics and demographical data were analyzed using SPSS 18.

4. Results

4.1. Descriptives

Frequencies for all three dimensions are summarized in Table 1. Analyses for dimension 1, *rule identification*, showed that a few participants learned the use of VOTAT to a certain degree during the first three items. Such learning or acquisition phases can only be observed if multiple items are used. However, if all items are considered, *rule identification* was largely constant throughout testing (see Table 2; SD = 0.06). Regarding dimension 2, *rule knowledge*, items with side effects or autoregressive processes (items 6–8) were much more difficult to understand than items without such effects (items 1–5) and thus, performance depended strongly on system structure. However, this classification did not fully account for *rule application*. Items were generally more difficult if participants had to control side effects or autoregressive processes (items 6–7) or items in which values of some variables had to be increased while others had to be decreased, respectively (items 2 and 4).

Internal consistencies as well as Rasch reliability estimates of MicroDYN were good to acceptable (Table 2). Not surprisingly, these estimates were, due to a *Multiple-Item-Approach,* somewhat lower than in other CPS scenarios. *One-Item-Testing* typically leads to dependencies of performance indicators likely to inflate internal consistencies. Cronbach's α of APM (α = 0.85) as well as participants' raw score distribution on APM (M = 25.67, s = 5.69) were comparable to the original scaling sample of university students (α = 0.82; M = 25.19, s = 5.25; Raven et al., 1998). The range of participants' GPA was restricted, indicating that

Table 1
Relative frequencies for the dimensions rule identification, rule knowledge and rule application (n = 222).

	Dimension 1: Rule identification		Dimension 2: Rule knowledge		Dimension 3: Rule application	
	0 no VOTAT	1 VOTAT	0 false	1 correct	0 false	1 correct
Item1	0.26	0.74	0.19	0.81	0.24	0.76
Item2	0.23	0.77	0.17	0.83	0.53	0.47
Item3	0.16	0.84	0.17	0.83	0.37	0.62
Item4	0.13	0.87	0.14	0.86	0.50	0.50
Item5	0.10	0.90	0.10	0.90	0.26	0.74
Item6	0.11	0.89	0.79	0.21	0.53	0.47
Item7	0.10	0.90	0.71	0.29	0.48	0.52
Item8	0.10	0.90	0.93	0.07	0.30	0.70

Note. VOTAT (Vary One Thing At A Time) describes use of the optimal strategy.

Table 2
Item statistics and reliability estimates for rule identification, rule knowledge and rule application (n = 222).

	Item statistics		Reliability estimates	
	M	*SD*	*Rasch*	α
Rule identification	0.85	0.06	0.82	0.86
Rule knowledge	0.60	0.34	0.85	0.73
Rule application	0.60	0.12	0.81	0.79

Note. M = mean; SD = standard deviation; Rasch = EA/PV reliability estimate within the Rasch model (1PL model); α = Cronbach's α; range for *rule identification*, *rule knowledge* and *rule application*: 0 to 1.

participants were mostly well above average performance (M = 1.7, s = 0.7; 1 = best performance, 6 = insufficient).

4.2. Measurement model for reasoning

To derive a measurement for reasoning, we divided APM scores in three parcels each consisting of 12 APM Set II-items. Using the item-to-construct balance recommended by Little, Cunningham, Shahar, and Widaman (2002), the highest three factor loadings were chosen as anchors of the parcels. Subsequently, we repeatedly added the three items with the next highest factor loadings to the anchors in inverted order, followed by the subsequent three items with highest factor loadings in normal order and so on. Mean difficulty of the three parcels did not differ significantly ($M_1 = 0.74$; $M_2 = 0.67$; $M_3 = 0.73$; $F_{2, 33} = 0.31$; $p > 0.05$).

4.3. Hypothesis 1: Measurement model of CPS

4.3.1. CFA

We ran a CFA to determine the internal structure of CPS. The assumed 3-dimensional model showed a good global model fit (Table 3), indicated by a Comparative Fit Index (CFI) and a Tucker Lewis Index (TLI) value above 0.95 and a Root Mean Square Error of Approximation (RMSEA) just within the limit of 0.06 recommended by Hu and Bentler (1999). However, Yu (2002) showed that RMSEA is too conservative in small samples.

Surprisingly, in the 3-dimensional model *rule identification* and *rule knowledge* were highly correlated on a latent level (r = 0.97). Thus, students who used VOTAT also drew appropriate conclusions, yielding in better *rule knowledge* scores. A descriptive analyses of the data showed that the probability to build a correct model without using VOTAT was 3.4% on average, excluding the first and easiest item which had a probability of 80%. Thus, the latent correlation between *rule identification* and *rule knowledge* based on empirical data was higher than theoretically assumed.

Concerning the internal structure of MicroDYN, a χ^2-difference test carried out subsequently (using Weighted Least Squares Mean and Variance adjusted — WLSMV estimator for ordinal variables, Muthén & Muthén, 2007b) showed that a more parsimonious 2-dimensional model with an aggregated facet of *rule knowledge* and *rule identification* on one factor and *rule application* on another factor did not fit significantly worse than the presumed 3-dimensional model ($\chi^2 = 0.821$; $df = 2$; $p > 0.05$), but better than a 1-dimensional model with all indicators combined on one factor ($\chi^2 = 17.299$; $df = 1$; $p < 0.001$). This indicated that, empirically, there was no difference between the facets *rule identification* and *rule knowledge*. Therefore, we decided to use only indicators of *rule knowledge* and not those of *rule identification,* because *rule knowledge* is more closely related to *rule application* in the process model (Kröner et al., 2005) as well as more frequently used in CPS literature as an indicator for generating information than *rule identification* (Funke, 2001; Kluge, 2008). It would also have been possible to use a 2-dimensional model with rule identification and rule knowledge combined under one factor and rule application under the other one. However, this model is less parsimonious (more parameters to be estimated) and the global model fit did not significantly increase.

Table 3
Goodness of Fit indices for measurement models including rule identification (RI), rule knowledge (RK) and rule application (RA) (n = 222).

MicroDYN Internal Structure	χ^2	*df*	p	χ^2/df	CFI	TLI	RMSEA
RI + RK + RA (3-dimensional)	82.777	46	0.001	1.80	0.989	0.991	0.060
RI & RK + RA (2-dimensional)	81.851	46	0.001	1.78	0.989	0.992	0.059
RI & RK & RA (1-dimensional)	101.449	46	0.001	2.20	0.983	0.987	0.074
RK & RA (1-dimensional)	78.003	41	0.001	1.90	0.964	0.971	0.064
RK + RA (2-dimensional)	61.661	41	0.020	1.50	0.980	0.984	0.048

Note. *df* = degrees of freedom; CFI = Comparative Fit Index; TLI = Tucker Lewis Index; RMSEA = Root Mean Square Error of Approximation; χ^2 and *df* are estimated by WLSMV. & = Facets constitute one dimension; + = Facets constitute separate dimensions. The final model is marked in bold.

Thus, for further analyses, the 2-dimensional model with only *rule knowledge* and *rule application* was used. This model fit was better than a g-factor model with *rule knowledge* and *rule application* combined (χ^2-difference test = 15.696, $df = 1$, $p < 0.001$), also showing a good global model fit (Table 3). The communalities ($h^2 = 0.36$–0.84 for *rule knowledge*; $h^2 = 0.08$–0.84 for *rule application*; see also Appendix B) were mostly well above the recommended level of 0.40 (Hair, Anderson, Tatham, & Black, 1998). Only item 6 showed a low communality on *rule application*, because it was the first item containing an autoregressive process, and participants underestimated the influence of this kind of effect while trying to reach a given target in the system.

4.3.2. IRT

After evaluating CFA results, we ran a multidimensional Rasch analysis on the 3-dimensional model, thereby forcing factor loadings to be equal, and changing the linear link function in CFA to a logarithmical one in IRT. Comparable to the results on CFA, *rule identification* and *rule knowledge* were highly correlated (r = 0.95), supporting the decision to focus on a 2-dimensional model. This model showed a significantly better fit than a 1-dimensional model including both facets ($\chi^2 = 34$; $df = 2$, $p < 0.001$), when a difference test of the final deviances as recommended by Wu, Adams, Wilson, and Haldane (2007) is used. Item fit indices (MNSQ) were within the endorsed boundaries from 0.75 to 1.33 (Bond & Fox, 2001), except for item 6 concerning *rule application*. Because item 6 fit well within *rule knowledge*, however, it was not excluded from further analyses.

Generally, both CFA and IRT results suggested that *rule application* can be separated from *rule knowledge* and *rule identification* while a distinction between the latter two could not be supported empirically. In summary, hypothesis 1 was only partially supported.

4.4. Hypothesis 2: Reasoning and CPS

We assumed that *rule knowledge* mediated the relationship between reasoning and *rule application*. In order to check mediation, it was expected that reasoning predicted *rule knowledge* and *rule application*, whereas prediction of *rule application* should no longer be significant if a direct path from *rule knowledge* to *rule application* was added.

Although a considerable amount of variance remained unexplained, reasoning predicted both facets as expected (*rule knowledge*: $\beta = 0.63$; $p < 0.001$; $R^2 = 0.39$; *rule application*: $\beta = 0.56$; $p < 0.001$; $R^2 = 0.31$), showing a good overall model fit (model (a) in Table 4). Thus, more intelligent persons performed better than less intelligent ones in *rule knowledge* and *rule application*.

However, if a direct path from *rule knowledge* to *rule application* was added (see path (c) in Fig. 4), the direct prediction of *rule application* by APM (path b) was no longer significant ($p = 0.52$), shown as an insignificant path (b) in Fig. 4. Consequently, more intelligent persons outperformed less intelligent ones in *rule application*, because they acquired more *rule knowledge* beforehand. Thus, learning *rule knowledge* is a prerequisite for *rule application*.

Results were unchanged if a 3-dimensional model including *rule identification* was used. Thus, Hypothesis 2 was supported.

4.5. Hypothesis 3: Predictive and incremental validity of CPS

We claimed that CPS predicted performance in GPA beyond reasoning. In order to test this assumption, first we checked predictive validity of each construct separately and then added all constructs combined in another model to test incremental validity (please note: stepwise latent regression is not supported by MPlus; Muthén & Muthén, 2007b). Reasoning significantly predicted GPA ($\beta = 0.35$, $p < 0.001$) and explained about 12% of variance in a bivariate latent regression showing a good model fit (model b in Table 4). If only CPS-facets were included in the analysis, *rule knowledge* predicted GPA ($\beta = 0.31$, $p < 0.001$) and explained about 10% of variance, whereas *rule application* had no influence on GPA. This model also fitted well (model (c) in Table 4). If reasoning and the CPS-facets were added simultaneously in a model (model (d) in Table 4), 18% of GPA-variance was explained, indicating that 6% of variance is additionally explained in comparison to the model with only reasoning as predictor of GPA (model b). However, the CPS facets and reasoning were correlated ($r_{APM/RA} = 0.56$; $r_{APM/RK} = 0.63$). Thus, covariances between reasoning and CPS might also have influenced the estimates of the path coefficient of CPS, so that the influence which is solely attributable to CPS is not evidently shown within this model. Thus, we decided to run another analysis and investigate incremental validity of CPS by using only one single model. Within this model (shown in Fig. 5), *rule knowledge* and *rule application* were regressed on reasoning. The residuals of this regression, RK_{res} *and* RA_{res}, as well as reasoning itself, were used to predict performance in GPA.

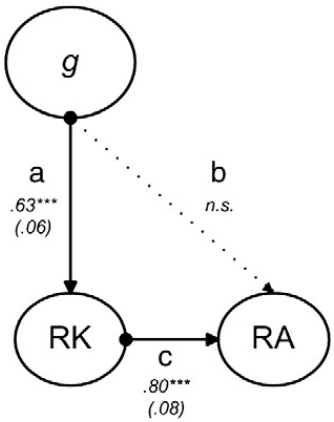

Fig. 4. Structural model including reasoning (g), MicroDYN rule knowledge (RK) and MicroDYN rule application (RA) (n = 222). Manifest variables are not depicted. *p<0.05; **p<0.01; ***p<0.001.

Results of this final model showed that reasoning predicted GPA, but the residual of *rule knowledge* RK_{res} explained additional variance in GPA beyond reasoning. RA_{res} yielded no significant path. Although this model is statistically identical to model (d), the significant path coefficient of RK_{res} showed incremental validity of CPS beyond reasoning more evidently, because RK_{res} and RA_{res} were modeled as independent from reasoning. In summary, RK_{res} involved aspects of CPS not measured by reasoning, but could predict performance in GPA beyond it. Thus, hypothesis 3 was supported.

5. Discussion

We extended criticisms by Kröner et al. (2005) on CPS research and tested a *Multiple-Item-Approach* to measure CPS. We claimed that (1) three different facets of CPS can be separated, (2) *rule knowledge* fully mediates the relationship between reasoning and *rule application* and (3) CPS shows

Table 4
Goodness of Fit indices for structural models including reasoning, CPS and GPA (n = 222).

	Hyp.	χ^2	*df*	p	χ^2/df	CFI	TLI	RMSEA
(a) Reasoning → CPS	2	79.554	50	0.005	1.59	0.967	0.979	0.052
(b) Reasoning → GPA	3	3.173	2	0.205	1.59	0.996	0.988	0.052
(c) CPS → GPA	3	69.181	46	0.015	1.50	0.977	0.982	0.048
(d) Reasoning & CPS → GPA	3	82.481	54	0.007	1.53	0.969	0.979	0.049

Note. *df* = degrees of freedom; CFI = Comparative Fit Index; TLI = Tucker Lewis Index; RMSEA = Root Mean Square Error of Approximation; χ^2 and *df* are estimated by WLSMV.

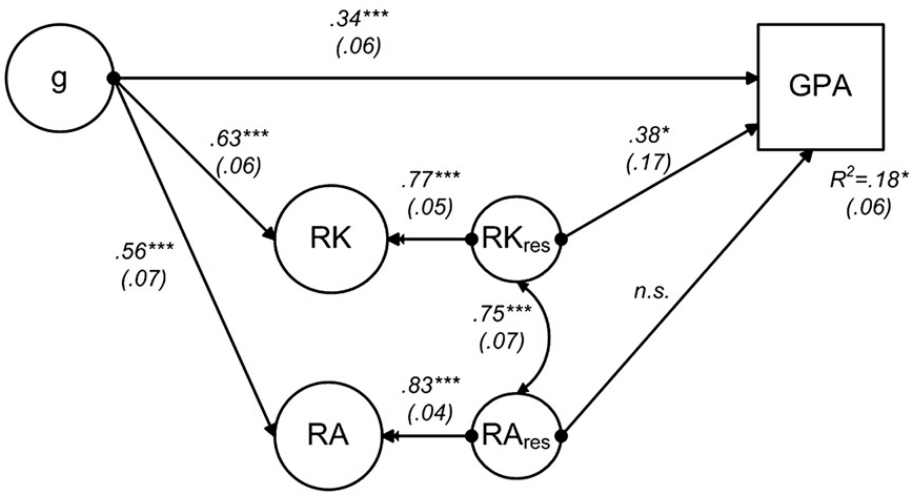

Fig. 5. Rule knowledge (RK) and rule application (RA) were regressed on reasoning (g). The residuals of this regression as well as reasoning were used to predict GPA. Manifest variables are not depicted. *p<0.05; **p<0.01; ***p<0.001.

incremental validity beyond reasoning. Generally, our findings suggest that CPS can be established as a valid construct and can be empirically separated from reasoning.

5.1. Ad (1) Internal structure

A three-dimensional model with the facets *rule identification*, *rule knowledge* and *rule application* was not supported (Hypothesis 1). Although *rule identification* and *rule knowledge* are theoretically distinguishable processes (Buchner, 1995), empirically there was no difference between them (r = 0.97). These findings differ considerably from results reported by Kröner et al. (2005), who conducted the only study including the measurement of *rule identification* as a CPS-facet in a process model of CPS. They reported a small, but significant path coefficient between both facets (r = 0.22) based on a sample of German high school students. However, theirs as well as our results might be influenced by methodological aspects. The low correlation between *rule identification* and *rule knowledge* found by Kröner et al. (2005) could be a result of assessing *rule knowledge* by forcing participants to predict values of a subsequent round and not only to assess mere knowledge about the system structure. Thus, *rule knowledge* is more similar to *rule application* (i.e., applying rules in order to reach goals), lowering correlations with *rule identification* (i.e., implementing appropriate strategies in order to identify relationships between variables). In contrast, in MicroDYN, the correlation may be overestimated, because the sample consisted of university students with above average cognitive performance. If these students used adequate strategies, they also drew correct conclusions leading to better performance in *rule knowledge*. The transfer from *rule identification* to *rule knowledge* may be more erroneous in a heterogeneous sample covering a broader range of cognitive ability. This may lead to an empirical separation of the two facets, which would either result if a considerable amount of students using VOTAT failed to draw correct conclusions about the systems' structure or students not using VOTAT succeeded in generating knowledge. Both were not the case in this study. Thus, it has to be tested if *rule identification* and *rule knowledge* can be empirically separated – as it is theoretically assumed – by using a representative sample and fully assessing participants' internal representation without forcing them to apply the rules at the same time.

However, results indicated that the operationalization of *rule identification* (VOTAT) was quite sufficient. According to the model depicted in Fig. 3, high *rule identification* scores should yield in good *rule knowledge* — and a strong relationship between both facets cannot be expected if indicators are not adequately chosen. Consequently, from a developmental point of view, it would be straightforward to teach an appropriate use of VOTAT to improve performance in *rule knowledge*. Within cognitive psychology, Chen and Klahr (1999) have made great endeavors to show that pupils can be trained to acquire VOTAT[1] in order to design unconfounded experiments (i.e., experiments that allow valid, causal inferences). In one experiment using hands-on material, pupils had to find out how different characteristics of a spring (e.g., length, width, and wire size) influenced how far it stretched. Trained pupils performed better than untrained ones in using VOTAT as well as in generalizing the knowledge gained across various contexts. Triona and Klahr (2003) and Klahr, Triona, and Williams (2007) extended this research and showed that using virtual material is also an effective method to train VOTAT within science education. Thus, domain unspecific CPS-skills assessed by MicroDYN and the skills taught in science education to discover physical laws experimentally seem to be very similar, so that the developmental implications of using MicroDYN as a training tool for domain-unspecific knowledge acquisition skills in school should be thoroughly investigated. We strongly encourage a comparison of these research fields in order to generalize contributions of CPS.

In summary, the ability of applying strategies – *rule identification* – can be theoretically distinguished from the ability of deriving *rule knowledge*. However, based on the results of

[1] Chen and Klahr (1999, p.1098) used the term *control of variables strategy* (CVS). CVS is a method for creating experiments in which a single contrast is made between experimental conditions and involves VOTAT.

this study, it is unclear if *rule identification* and *rule knowledge* can be empirically separated, although VOTAT was an appropriate operationalization of *rule identification* for the items used within linear structural equation systems. If items based on other approaches are used, other indicators for rule identification may be more appropriate. Finally, data suggests a clear distinction between *rule knowledge* and *rule application* also supported by previous research, even though within *One-Item-Testing* (Beckmann & Guthke, 1995; Funke, 2001; Kröner et al., 2005).

5.2. Ad (2) CPS and reasoning

In a bivariate model, reasoning predicted both *rule knowledge* and *rule application*. However, 60% of variance in *rule knowledge* and 69% of variance in *rule application* remained unexplained, suggesting that parts of the facets are determined by other constructs than reasoning. Furthermore, in a process model of CPS, *rule knowledge* mediated the relationship between reasoning and *rule application*, whereas the direct influence of reasoning was not significant. The insignificant direct path from reasoning to *rule application* indicated that more intelligent persons showed better rule application performance than less intelligent ones not directly because of their intelligence, but because they used their abilities to acquire more *rule knowledge* beforehand.

These results are contrary to Kröner et al. (2005), who reported a direct prediction of *rule application* by reasoning. This indicates that a lack of *rule knowledge* could be partly compensated by reasoning abilities (p. 364), which was not the case in the present study, although participants were allowed to use the model showing the correct system structure. However, their result might be due to *rule application* measured as one-step control round without giving feedback. Thus, the ability to counteract unwanted developments based on dynamic system changes as well as using feedback is not assessed and important cognitive operations allocated to CPS tasks like evaluating ones own decisions and adapting action plans are not measured (Funke, 2001). Consequently, *rule application* depends significantly more on reasoning (Kröner et al., 2005).

In summary, reasoning is directly related to the CPS-process of generating knowledge. However, a considerable amount of CPS variance remained unexplained. In order to actively reach certain targets in a system, sufficient *rule knowledge* is a prerequisite for *rule application*.

5.3. Ad (3) Construct validity

Using data from the German national extension study in PISA 2000, Wirth et al. (2005) showed that performance in CPS (measured by *Space Shuttle*) is correlated with PISA-test performance in school subjects like maths, reading and sciences ($r = 0.25–0.48$). In the present study, this finding was extended by showing for the first time that CPS predicts performance in GPA even beyond reasoning. This result shows the potential of CPS as a predictor of cognitive performance. It also emphasizes that it is important to measure different problem solving facets, and not *rule application* exclusively as indicator of CPS performance as occasionally has been done (Gonzalez, Thomas, & Vanyukov, 2005), because residual parts of rule knowledge RK_{res}, explained variance in GPA beyond reasoning while RA_{res} did not. Thus, *rule knowledge* – the ability to draw conclusions in order to generate knowledge – was more closely connected to GPA than *rule application* — the ability to use knowledge in order to control a system. This is not surprising, because acquiring knowledge is more frequently demanded in school subjects than using information in order to actively control a system (Lynch & Macbeth, 1998; OECD, 2009). For *rule application*, however, criteria for assessing predictive validity are yet to be found. For instance, measuring employees' abilities in handling machines in a manufactory might be considered, because workers are used to getting feedback about actions immediately (e.g., a machine stops working) and have to incorporate this information in order to actively control the machine (e.g., take steps to repair it).

Several shortcomings in this study need consideration: (1) The non-representative sample entails a reduced generalizability (Brennan, 1983). A homogenous sample may lead to reduced correlations between facets of CPS, which in turn may result in more factorial solutions in SEM. Consequently, the 2-dimensional model of CPS has to be regarded as a tentative result. Additionally, a homogenous sample may lead to lower correlations between reasoning and CPS (Rost, 2009). However, APM was designed for assessing performance in samples with above average performance (Raven, 1958). Participants' raw score distribution in this study was comparable to the original scaling sample of university students (Raven et al., 1998) and variance in APM and also in MicroDYN was sufficient. The selection process of the university itself considered only students' GPA. Thus, variance on GPA was restricted, but even for this restricted criterion CPS showed incremental validity beyond reasoning. Furthermore, in studies using more representative samples, residual variances of CPS facets like *rule application* also remained unexplained by reasoning (93% of unexplained variance in Bühner et al., 2008; 64% of unexplained variance in Kröner et al., 2005) indicating the potential increment of CPS beyond reasoning. Nevertheless, an extension of research using a more heterogeneous sample with a broad range of achievement potential is needed.

(2) Moreover, it could be remarked that by measuring reasoning we tested a rather narrow aspect of intelligence. However, reasoning is considered to be at the core of intelligence (Carroll, 1993) and the APM is one of the most frequently used as well as broadly accepted measurement devices in studies investigating the relationship between CPS and intelligence (Gonzalez, Thomas, & Vanyukov, 2005; Goode & Beckmann, 2011). Nevertheless, in a follow-up experiment, a broader operationalization of intelligence may be useful. The question of which measurement device of intelligence is preferable is closely related to the question of how CPS and intelligence are related on a conceptual level. Within Carrolls' three stratum theory of intelligence (1993, 2003), an overarching ability factor is assumed on the highest level (stratum 3), which explains correlations between eight mental abilities located at the second stratum, namely fluid and crystallized intelligence, detection speed, visual or auditory perception, general memory and learning, retrieval ability, cognitive speediness and processing speed. These factors explain performance in 64 specific, but correlated abilities (located on stratum 1). Due to empirical results of the last

two decades which have reported correlations between intelligence and reliable CPS tests, researchers in the field would probably agree that performance on CPS tasks is influenced by general mental ability (stratum 3). But how exactly is CPS connected to factors on stratum 2 that are usually measured in classical intelligence tests? Is CPS a part of the eight strata mentioned by Carroll (1993), or is it an ability that cannot be subsumed within stratum 2? Considering our results on incremental validity, CPS ability may constitute at least some aspects of general mental ability divergent from reasoning. This assumption is also supported by Danner, Hagemann, Schankin, Hager, and Funke (2011), who showed that CPS (measured by *Space Shuttle* and *Tailorshop*) predicted supervisors' ratings even beyond reasoning (measured by subscale processing capacity of the *Berlin Intelligence Structure Test* and by *Advanced Progressive Matrices*, APM). Concerning another factor on stratum 2, working memory, Bühner et al. (2008) showed that controlling for it reduced all paths between intelligence (measured by figural subtests of *Intelligence Structure Test 2000 R*, Amthauer et al., 2001), rule knowledge, and rule application (both measured by *MultiFlux*) to insignificance. Thus, they concluded that working memory is important for computer-simulated problem-solving scenarios. However, regarding *rule application*, working memory is more necessary if problem solvers have only one control round in order to achieve goals as realized within MultiFlux, because they have to incorporate effects of multiple variables (i.e., controls) simultaneously. Contrarily, if CPS tasks consist of multiple control rounds, problem solvers may use the feedback given, which is less demanding for working memory. Consequently, the influence of working memory on CPS tasks may at least partly depend on the operationalization used.

Empirical findings on the relationship of CPS to other factors mentioned on the second stratum by Carroll (2003) are yet to be found. However, all these factors are measured by static tasks that do not assess participants' ability to actively generate and integrate information (Funke, 2001; Greiff, in press), although tests exist, which include feedback that participants may use in order to adjust behavior. These tests are commonly aimed to measure learning ability (e.g., in reasoning tasks) as captured in the facet long-term storage and retrieval (Glr; Carroll, 2003). Participants may either be allowed to use feedback to answer future questions (e.g., *Snijders-Oomen non-verbal intelligence test — SON-R*, Tellegen, Laros, & Petermann, 2007) or to answer the very same question once again (e.g., *Adaptive Computer supported Intelligence Learning test battery — ACIL*; Guthke, Beckmann, Stein, Rittner, & Vahle, 1995). The latter approach is most similar to CPS. However, Glr is often not included in the "core set" of traditional intelligence tests and the tasks used do not contain several characteristics of complex problems that are assessed in MicroDYN, e.g., connectedness of variables or intransparency. These characteristics require from the problem solver to actively generate information, to build a mental model and to reach certain goals. Nevertheless, a comparison of MicroDYN and tests including feedback should be conducted in order to provide more information on how closely CPS and learning tests are related.

In summary, as CPS captures dynamic and interactive aspects, it can be assumed that it constitutes a part of general mental ability usually not assessed by classical intelligence tests covering the second stratum factors of Carroll (2003). Research on CPS at a sound psychometrical level started only about a decade ago and, thus, adequate instruments for CPS have not been available for Carrolls' analyses involving factor analysis for a huge amount of studies that were done before the 90s.

Independently of where exactly CPS should be located within Carrolls' 3 strata, as a construct it contributes considerably to the prediction of human performance in dealing with unknown situations that people encounter almost anywhere in daily life — a fact that has been partially denied by researchers. It should not be.

Acknowledgments

This research was funded by a grant of the German Research Foundation (DFG Fu 173/14-1). We gratefully thank Andreas Fischer and Daniel Danner for their comments.

Appendix A

The 8 items in this study were mainly varied regarding two system attributes proved to have the most influence on item difficulty (see Greiff, in press): the number of effects between the variables and the quality of effects (i.e., with or without side effects/autoregressive processes). All other variables are held constant (e.g., strength of effects, number of inputs necessary for optimal solutions, etc.).

	Linear structural equations	*System size*	*Effects*
Item 1	$Xt+1=1*Xt+0*At+2*Bt$ $Yt+1=1*Yt+0*At+2*Bt$	2×2-System	Only direct
Item 2	$Xt+1=1*Xt+2*At+2*Bt+0*Ct$ $Yt+1=1*Yt+0*At+0*Bt+2*Ct$	2×3-System	Only direct
Item 3	$Xt+1=1*Xt+2*At+2*Bt+0*Ct$ $Yt+1=1*Yt+0*At+2*Bt+0*Ct$ $Zt+1=1*Zt+0*At+0*Bt+2*Ct$	3×3-System	Only direct
Item 4	$Xt+1=1*Xt+2*At+0**Bt+0*Ct$ $Yt+1=1*Yt+0*At+2*Bt+2*Ct$ $Zt+1=1*Zt+0*At+0*Bt+2*Ct$	3×3-System	Only direct
Item 5	$Xt+1=1*Xt+2*At+0*Bt+2*Ct$ $Yt+1=1*Yt+0*At+2*Bt+0*Ct$ $Zt+1=1*Zt+0*At+0*Bt+2*Ct$	3×3-System	Only direct
Item 6	$Xt+1=1.33*Xt+2*At+0*Bt+0**Ct$ $Yt+1=1*Yt+0*At+0*Bt+2*Ct$	2×3-System	Direct and indirect
Item 7	$Xt+1=1*Xt+0.2*Yt+2*At+2*Bt+0*Ct$ $Yt+1=1*Yt+0*At+0*Bt+0*Ct$	2×3-System	Direct and indirect
Item 8	$Xt+1=1*Xt+2*At+0*Bt+0*Ct$ $Yt+1=1*Yt+2*At+0*Bt+0*Ct$ $Zt+1=1.33*Zt+0*At+0*Bt+2*Ct$	3×3-System	Direct and indirect

Note. X_t, Y_t, and Z_t denote the values of the output variables, and A_t, B_t, and C_t denote the values of the input variables during the present trial, while X_{t+1}, Y_{t+1}, Z_{t+1} denote the values of the output variables in the subsequent trial.

Appendix B

Factor loadings and communalities for *rule identification*, *rule knowledge* and *rule application* (n = 222).

	Rule identification		Rule knowledge		Rule application	
	Factor loading	h^2	*Factor loading*	h^2	*Factor loading*	h^2
Item 1	0.70	0.49	0.73	0.53	0.50	0.25
Item 2	0.90	0.81	0.74	0.55	0.84	0.71
Item 3	0.92	0.85	0.88	0.77	0.83	0.69
Item 4	0.99	0.98	0.91	0.83	0.90	0.81
Item 5	0.99	0.98	0.94	0.88	0.92	0.85
Item 6	0.92	0.85	0.63	0.40	0.26	0.07
Item 7	0.95	0.90	0.70	0.49	0.68	0.46
Item 8	0.95	0.90	0.46	0.21	0.75	0.56

Note. All loadings are significant at $p<0.01$.

References

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *Intelligenz-Struktur-Test 2000 R [Intelligence Structure Test 2000 R].* Göttingen: Hogrefe.

Babcock, R. L. (2002). Analysis of age differences in types of errors on the Raven's advanced progressive matrices. *Intelligence*, *30*, 485–503.

Beckmann, J. F. (1994). *Lernen und komplexes Problemlösen: Ein Beitrag zur Konstruktvalidierung von Lerntests [Learning and complex problem solving: A contribution to the construct validation of tests of learning potential].* Bonn, Germany: Holos.

Beckmann, J. F., & Guthke, J. (1995). Complex problem solving, intelligence, and learning ability. In P. A. Frensch, & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 177–200). Hillsdale, NJ: Erlbaum.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Erlbaum.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City, IA: American College Testing.

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.

Bühner, M., Kröner, S., & Ziegler, M. (2008). Working memory, visual–spatial intelligence and their relationship to problem-solving. *Intelligence*, *36* (4), 672–680.

Burns, B. D., & Vollmeyer, R. (2002). Goal specificity effects on hypothesis testing in problem solving. *Quarterly Journal of Experimental Psychology*, *55A*, 241–261.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge: Cambridge University Press.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports g and about ten broad factors. In H. Nyborg (Ed.), *Thescientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 5–21). Amsterdam, NL: Pergamon.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the Control of Variables Strategy. *Child Development*, *70*(5), 1098–1120.

Danner, D., Hagemann, D., Schankin, A., Hager, M., & Funke, J. (2011). Beyond IQ. A latent state trait analysis of general intelligence, dynamic decision making, and implicit learning. *Intelligence*, *39*(5), 323–334.

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

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Eysenck, H. J. (2000). *Intelligence: A new look.* New Brunswick, NJ: USA, Transaction.

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

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

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

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

Gonzalez, C., Vanyukov, P., & Martin, M. K. (2005). The use of microworlds to study dynamic decision making. *Computers in Human Behavior*, *21*(2), 273–286.

Goode, N., & Beckmann, J. (2011). You need to know: There is a causal relationship between structural knowledge and control performance in complex problem solving tasks. *Intelligence*, *38*, 345–552.

Greiff, S. (in press). *Individualdiagnostik der Problemlösefähigkeit.* [Diagnostics of problem solving ability on an individual level]. Münster: Waxmann.

Greiff, S., & Funke, J. (2010). Systematische Erforschung komplexer Problemlösefähigkeit anhand minimal komplexer Systeme [Some systematic research on complex problem solving ability by means of minimal complex systems]. *Zeitschrift für Pädagogik*, *56*, 216–227.

Guthke, J., Beckmann, J. F., Stein, H., Rittner, S., & Vahle, H. (1995). *Adaptive Computergestützte Intelligenz-Lerntestbatterie (ACIL) [Adaptive computer supported intelligence learning test battery].* : Mödlingen: Schuhfried.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. (1998). *Multivariate data analysis.* Upper Saddle River, NJ: Prentice Hall.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Jäger, A. O., Süß, H. M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test, Form 4 [Berlin Intelligence Structure Test].* Göttingen, Germany: Hogrefe.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT, US: Praeger Publishers/Greenwood Publishing Group.

Jensen, A. R. (1998). The g factor and the design of education. In R. J. Sternberg, & W. M. Williams (Eds.), *Intelligence, instruction, and assessment. Theory into practice* (pp. 111–131). Mahwah, NJ, USA: Erlbaum.

Joslyn, S., & Hunt, E. (1998). Evaluating individual differences in response to time–pressure situations. *Journal of Experimental Psychology*, *4*, 16–43.

Klahr, D., Triona, L. M., & Williams, C. (2007). Hands on what? The relative effectiveness of physical versus virtual materials in an engineering desing project by middle school children. *Journal of Research in Science Teaching*, *44*, 183–203.

Klieme, E., Funke, J., Leutner, D., Reimann, P., & Wirth, J. (2001). Problemlösen als fächerübergreifende Kompetenz. Konzeption und erste Resultate aus einer Schulleistungsstudie [Problem solving as crosscurricular competency. Conception and first results out of a school performance study]. *Zeitschrift für Pädagogik*, *47*, 179–200.

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

Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence*, *33*(4), 347–368.

Leighton, J. P. (2004). Defining and describing reason. In J. P. Leighton, & R. J. Sternberg (Eds.), *The Nature of Reasoning* (pp. 3–11). Cambridge: Cambridge University Press.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, *9*(2), 151–173.

Lynch, M., & Macbeth, D. (1998). Demonstrating physics lessons. In J. G. Greeno, & S. V. Goldman (Eds.), *Thinking practices in mathematics and science learning* (pp. 269–298). Hillsdale, NJ: Erlbaum.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence*, *7*, 107–127.

Muthén, B. O., & Muthén, L. K. (2007). *MPlus.* Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2007). *MPlus user's guide.* Los Angeles, CA: Muthén & Muthén.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77–101.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H. M. (2005). Working memory and intelligence — Their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*, 61–65.

OECD (2004). *Problem solving for tomorrow's world.* First measures of cross-curricular competencies from PISA 2003. Paris: OECD.

OECD (2009). *PISA 2009 assessment framework — Key competencies in reading, mathematics and science.* Paris: OECD.

Putz-Osterloh, W. (1981). Über die Beziehung zwischen Testintelligenz und Problemlöseerfolg [On the relationship between test intelligence and success in problem solving]. *Zeitschrift für Psychologie*, *189*, 79–100.

Raven, J. C. (1958). *Advanced progressive matrices* (2nd ed.). London: Lewis.

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

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's progressive matrices and vocabulary scales: Section 4.* The advanced progressive matrices. San Antonio, TX: Harcourt Assessment.

Rigas, G., Carling, E., & Brehmer, B. (2002). Reliability and validity of performance measures in microworlds. *Intelligence*, *30*, 463–480.

Rost, D. H. (2009). *Intelligenz: Fakten und Mythen [Intelligence: Facts and myths]* (1. Aufl. ed.). Weinheim: Beltz PVU.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology*, *86*(1), 162–173.

Sternberg, R. J., Conway, B. E., Ketron, J. L., & Bernstein, M. (1981). People's conceptions of intelligence. *Journal of Personality and Social Psychology*, *41*(1), 37–55.

Sternberg, R. J., Grigorenko, E. L., & Bundy, D. A. (2001). The predictive value of IQ. *Merrill-Palmer Quarterly: Journal of Developmental Psychology*, *47*(1), 1–41.

Süß, H. -M. (1996). *Intelligenz, Wissen und Problemlösen: Kognitive Voraussetzungen für erfolgreiches Handeln bei computersimulierten Problemen [Intelligence, knowledge, and problem solving: Cognitive prerequisites for success in problem solving with computer-simulated problems].* Göttingen: Hogrefe.

Süß, H. -M., Kersting, M., & Oberauer, K. (1993). Zur Vorhersage von Steuerungsleistungen an computersimulierten Systemen durch Wissen und Intelligenz [The prediction of control performance in computer based systems by knowledge and intelligence]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *14*, 189–203.

Tellegen, P. J., Laros, J. A., & Petermann, F. (2007). *Non-verbaler Intelligenztest: SON-R 2 1/2–7.Test manual mit deutscher Normierung und Validierung [Non-verbal intelligence test: SON-R].* Wien: 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.

Vollmeyer, R., & Rheinberg, F. (1999). Motivation and metacognition when learning a complex system. *European Journal of Psychology of Education*, *14*, 541–554.

Weiß, R. H. (2006). *Grundintelligenztest Skala 2 — Revision CFT 20-R [Culture fair intelligence test scale 2 — Revision].* Göttingen: Hogrefe.

Wiley, J., Jarosz, A. F., Cushen, P. J., & Colflesh, G. J. H. (2011). New rule use drives the relation between working memory capacity and Raven's Advanced Progressive Matrices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 256–263.

Wirth, J. (2004). *Selbstregulation von Lernprozessen [Self-regulation of learning processes].* Münster: Waxmann.

Wirth, J., & Leutner, D. (2008). Self-regulated learning as a competence. Implications of theoretical models for assessment methods. *Journal of Psychology*, *216*, 102–110.

Wirth, J., Leutner, D., & Klieme, E. (2005). Problemlösekompetenz – Ökonomisch und zugleich differenziert erfassbar? In E. Klieme, D. Leutner, & J. Wirth (Eds.), *Problemlösekompetenz von Schülerinnen und Schülern* (pp. 7–20). *Problem solving competence for pupils* (pp. 7–20). Wiesbaden: VS Verlag für Sozialwissenschaften.

Wittmann, W., & Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science*, *21* 393–40.

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

Wu, M. L., Adams, R. J., & Haldane, S. A. (2005). *ConQuest (Version 3.1).* Berkeley, CA: University of California.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest 2.0: Generalised item response modelling software [computer program manual].* Camberwell, Australia: Australian Council for Educational Research.

Yu, C. -Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Los Angeles, CA: University of California.